CONFOUNDING FACTOR CORRECTION FOR ACCURATE EXPRESSION QUANTITATIVE TRAIT LOCI DISCOVERY

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jin Hyun Ju

May 2017

CONFOUNDING FACTOR CORRECTION FOR ACCURATE EXPRESSION QUANTITATIVE TRAIT LOCI DISCOVERY

Jin Hyun Ju, Ph.D.

Cornell University 2017

Expression quantitative trait loci (eQTL) have become an attractive research topic in the past decade assisted by the technical advances in next-generation sequencing (NGS) and high-throughput gene expression measurements. eQTL discoveries provided researchers with new insights into genetic regulatory mechanisms, and are crucial in establishing functional links in genome-wide association study (GWAS) results. A powerful aspect of these studies are that the simultaneous genome wide measurements of gene expression values and sequence variants make it possible to detect associations independent of prior knowledge. However, the high dimensionality of the data also creates multiple challenges in the analysis process. Population structure in genotype data can induce significant inflation in the results leading to false positive findings, and confounding factors in gene expression measurements, such as technical batch effects and environmental differences, can lower the detection power of small genetic effects.

The focus of this thesis is on the challenges in analyzing high-dimensional gene expression data to increase the accuracy in eQTL discovery. A central problem in developing confounding factor correction methods for eQTL analysis is to account for non-genetic confounding factors, while preventing broad impact genetic effects of being modeled as non-genetic variation. To address this issue, we developed a novel method CONFETI: CONfounding Factor Estima-

tion Through Independent component analysis. CONFETI is based on a linear mixed model framework and uses independent component analysis (ICA) to estimate statistically independent generative sources from the observed gene expression profiles. Candidate genetic effects are excluded from the correction to maximize the discovery of broad impact eQTL, using the estimated independent components. We evaluated our framework by comparing the performance to other published confounding factor correction methods using both simulated and real human data.

In the analysis of simulated data, we show that CONFETI most accurately recovered simulated eQTL results in the presence of confounding factors by distinguishing genetic effects from non-genetic variance. We then analyzed matched twin pair datasets from the Multiple Tissue Human Expression Resource (MuTHER) consortium and datasets consisting of similar tissue pairs from the Genotype-Tissue Expression (GTEx) consortium. To assess the performance of each method in human data, we investigated the replication of *cis* and *trans*-eQTL identified in each dataset. We found that accounting for confounding factors greatly increased both the number of identified *cis*-eQTL in each dataset, and replicating *cis*-eQTL between twin pairs and similar tissue types. The number of identified *trans*-eQTL increased as well, however, most of the findings were specific to each dataset and the replication rate remained significantly lower compared to *cis*-eQTL. While the use of confounding factor correction methods increased the power of the analysis, we found little difference in identifying replicating *cis* and *trans*-eQTL in human data by removing candidate genetic effects prior to correction.

BIOGRAPHICAL SKETCH

Driven by the curiosity to better understand the common language of life, DNA, Jin Hyun Ju began studying biotechnology in 2006 at Yonsei University, Seoul, South Korea. He had to put his life as a student briefly on hold for two years to fulfill his mandatory duty in the Korean military from 2008 to 2010. After completing his duties as a Korean citizen, he had the opportunity to spend a year abroad at the University of California, San Diego. Upon the completion of his bachelor's degree in 2012, he joined the Physiology, Biophysics, and Systems Biology graduate program at the Weill Cornell Graduate School. There he met his mentor Jason Mezey and worked on developing statistical methods for studying expression quantitative trait loci. Jin Hyun will continue his research in genomics as a Bioinformatics Scientist in the Oncology Business Unit at Illumina.

To my family and anyone who will actually read this.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my mentor Jason Mezey, who gave me the opportunity to grow by pointing me in the right direction while being the most patient and understanding person I have ever met. Your method of teaching complex concepts in a very simple and intuitive manner not only expanded my knowledge in statistical genomics, but it also changed the way I approach new subjects and communicate with others.

I would also like to thank my committee members Olivier Elemento, Gunnar Rätsch, Joseph Pickrell, and Ekta Khurana for their valuable insights and helpful comments that helped my project develop over the years. I learned a lot from everyone of you in becoming an independent researcher and critical thinker.

Our lab's post-doc-in-charge Sushila, without whom my research would not have been possible, deserves special thanks for her selfless dedication to help others and great but somewhat weird sense of humor. I would also like to thank our lab members Monica, Francisco, Sarah, Mark, Abishek, Zijun, Afrah, Juan, Gabe, Thomas, Yuxin for being wonderful colleagues and friends.

Thank you Sungou, my role model, for introducing me to the field of bioinformatics and all the helpful advices which guided me through graduate school.

For mental and beer support and keeping me within the boundaries of sanity, I would like to thank all my friends I met in the great city of New York. Special thanks to Charlie, Litsa, Rudy, Vicky, James, Jenny, Alec and Davinder for all the fun times at Koliba et al., Johnathan for a matching level of pessimism and constructive sarcasm, Kjong for being my spiritual teacher in statistics, Adrian and Jana for all the fun discussions and excuses to devour chocolate ice cream, Friedie and Anatol for brain teasing game nights, Jason B. and Luce for both being great teachers and friends.

Thanks to my wonderful parents for showing me nothing but love and support, and to my parents-in-law for always looking out for me. Finally, I would like to thank InSun who was by my side from day one of my graduate school journey as a loving wife, best friend, and an objectively awesome person. You have always put a smile on me and filled my life with happiness. I love you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Technical advances in measuring genome-wide genetic variation, including single nucledotie polymorphism (SNP) arrays and next-generation sequencing (NGS) techniques, transformed the field of genetics and introduced us to the genomics era. We moved from investigating the genome with linkage analysis and candidate-gene based methods to hypothesis generating Genome-wide Association Studies (GWAS) [1]. The concept of GWAS was formulated after the discovery of linkage disequilibrium (LD) blocks, which are local correlation structures in the genome that are inherited together. While it is difficult to measure the exact causal polymorphism without the use of whole genome sequencing, LD blocks led to the revelation that it is possible to proxy the causal polymorphism by having a subset of marker SNPs throughout the genome [2]. In a typical GWAS, the association between genome-wide sequence variations and a trait of interest is tested. Ever since the first GWAS in 2005 [3], GWAS results have been at the forefront of identifying genomic regions associated with specific traits of interest, such as common diseases and phenotypes including height and body mass index (BMI). As of 2016, the GWAS catalog [4] contained 2,650 study results and 29,954 unique SNP-trait associations for various diseases and complex traits. In addition to prioritizing genes or genomic regions in studying specific traits, GWAS results can be used to reveal valuable insights related to the genetic architecture by estimating the heritability and effect size, and by identifying genetic interactions such as epistasis or pleiotropy [5].

However, despite the vast amount of accumulated knowledge generated by GWAS results, little is known about the underlying biological mechanisms in these findings [6]. Part of the reason stems from the fact that most GWAS findings fall within non-coding regions of the genome, making it difficult to draw direct conclusions of their biological impact [7]. Additionally, differences in gene regulation between tissues and cell types make it challenging to investigate the biological pathways relevant to the genomic regions of interest. One way to address this problem is to integrate information from GWAS results with that of gene regulatory regions identified by expression Quantitative Trait Loci (eQTL) analyses.

1.1 Expression Quantitative Trait Loci

eQTL are genomic sequence variants that influence the expression level of single or multiple genes [8]. These can be identified by testing the association between measured sequence variants, generally SNPs, and gene expression levels measured by microarray or RNA-sequencing (Figure 1.1). Basically, eQTL can be considered as multiple GWAS analyses where the trait of interests are gene expression levels. The current genome-wide picture of the genetics of gene expression in humans has been driven by eQTL studies in various populations and cell types [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. Findings in eQTL studies are routinely leveraged to identify candidate disease risk loci within regions associated with complex diseases in GWAS. This process is based on the assumption that when an eQTL co-locates with a locus identified in a GWAS, the same allelic variants are impacting both gene expression and disease

risk [23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45].

Genome-wide eQTL discovery has also provided a foundation for inferences about biological systems. For example, eQTL are used within data aggregation methods to annotate the functional or fitness impacts of polymorphisms [46], which in turn is a main component of systems biology models of pathways and cellular processes [47,48,49,50,51]. eQTL are also used for network modeling, in large part because eQTL can be used to model a directed impact on gene expression, which in turn can be leveraged to infer other directed network relationships among expressed genes [52,53,54,55].

Such eQTL discovery approaches have led to a number of generalizations about the genetics of gene expression and regulation at genome-wide scales [11, 56]. These include observations that the majority of genes in the genome can be impacted by an eQTL [57], that eQTL proximal to the regulated gene have significantly larger effect sizes than distant eQTL [18,30,58], and that eQTL can have tissue specific impacts on an expressed gene [34,59].

1.1.1 *cis*-eQTL and *trans*-eQTL

eQTL can be broadly classified into *cis*-eQTL and *trans*-eQTL based on the distance between the genotype and associated gene location. Albeit varying principles, eQTL are generally considered *cis* if the genotype and the associated gene are located on the same chromosome within a 1Mb distance, and *trans* other-

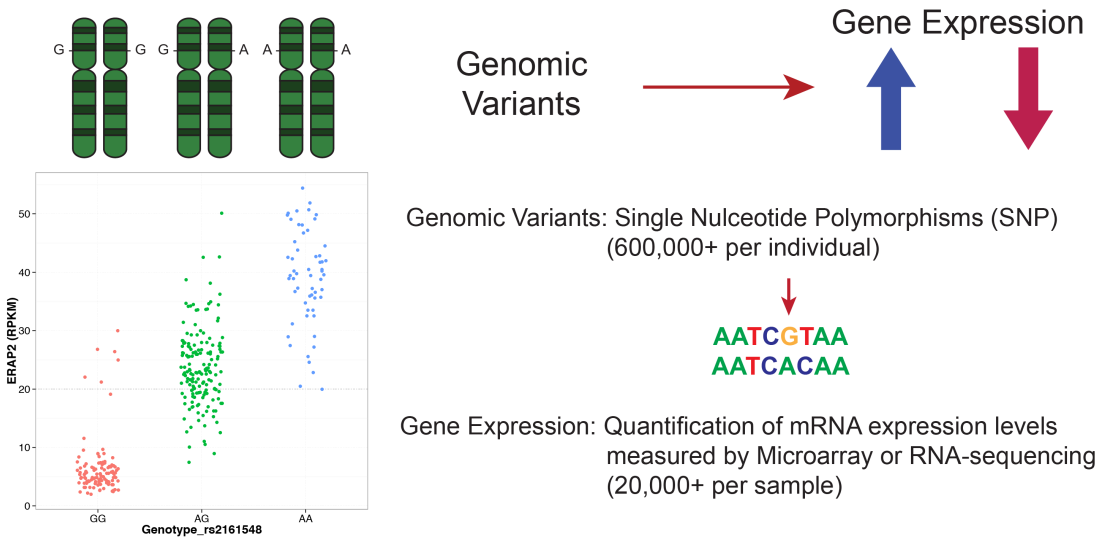


Figure 1.1: Illustration of eQTL

eQTL are sequence variants in the genome that are associated with expression level changes of specific genes. A typical eQTL analysis tests the association between approximately 600,000 genotypes and 20,000 gene expression measurements.

wise [18, 56]. Other classification methods have been proposed to account for allele specific effects [8, 37]. However, this thesis will only consider the distance based categorization.

Due to the large number of possible genotype-expression variable pair comparisons in human eQTL studies, which can range from $10^9 - 10^{10}$ for array based studies [15] and $10^{11} - 10^{12}$ for data collected by next-generation sequencing technologies [20], it is common to reduce the multiple testing burden by only considering a subset of genotype-expression pairs. Therefore, many studies have primarily focused on identifying *cis*-eQTL by only testing genotypes and phenotypes that are within a certain distance [11, 15, 17, 22]. A consequence of such strategies is that well defined characteristics for *cis*-eQTL have been established, such as enrichment near transcription start sites and large effect size,

but relatively little is known about *trans*-eQTL. While more recently *trans*-eQTL also came into focus [18,21,30,60], the number of identified *trans*-eQTL remain modest and evidence for replication varies [61]. It is hypothesized that the detection and replication of *trans*-eQTL is difficult partly due to their tissue or condition specific functionalities [59,62].

1.1.2 Broad Impact eQTL

When considering associations with complex diseases, eQTL that affect many genes have been hypothesized to have effects beyond the transcriptome and are therefore good candidates for investigating downstream disease phenotypes [63]. Such broad impact eQTL [21], variously referred to as eQTL hotspots [8], master regulators [60], *trans*-regulators [64], and *trans*-eQTL networks [65] could result from either hotspots of multiple co-located eQTL [8,66] or from the pleiotropic effects of a single eQTL genotype [37].

For studies that leverage eQTL as a foundation for network modeling or for identifying candidate disease risk loci, eQTL that are associated with multiple genes can be particularly valuable. Additionally, the value of such broad impact eQTL is clear for directed network modeling, since the network inference depends on tracking the impact of eQTL through multiple genes [67,68,69,70,71]. Broad impact eQTL have regularly been observed in model organisms such as yeast [72,73,74] and mice [75], but have been reported less frequently and in smaller numbers in human eQTL studies [8]. Based on observations that broad impact eQTL are expected to primarily affect *trans*-genes, statistical power has

been suggested as a possible reason for the relatively lower reporting of broad impact eQTL in humans, since *trans*-eQTL tend to have relatively weak associations in humans compared to model organisms [8]. Additionally, methods that account for unknown confounding variance have the potential to incorrectly explain away the effect of broad impact eQTL [76,77].

1.2 Challenges in eQTL Analyses

1.2.1 Lack of Standardization

The rapid advance in genotyping and gene expression measurement technologies have undoubtedly benefited the field of human genetics through the findings of eQTL studies. However, since the advancement in technology is still an ongoing process, for many eQTL studies the platforms for measuring gene expression and genotyping are different making it difficult to combined multiple datasets. It is common that the set of measured SNPs are different and that missing SNPs need to be imputed based on the available information. For gene expression measurements, microarrays generally have fewer genes measured compared to RNA-seq since they require capture probes, and both techniques are influenced by different technical artifacts. Finally, most of the analysis methods are unique to each study making it difficult to directly compare the results of multiple studies. Therefore, comparing results between different datasets are largely achieved by re-analyzing entire datasets with a centralized pipeline.

1.2.2 Population Structure

To detect eQTL with relatively small genetic effect sizes, a large number of samples are required for an association study. Therefore, it is common for studies with large scale datasets to have samples originating from multiple populations [15, 20, 22, 78]. In analyzing these samples consisting of different populations, accounting for population structure is a critical step to control the number of false positive findings. The presence of broad genetic similarity between samples can be problematic because an imbalance in the measured expression levels between populations could lead to all SNPs that are correlated to the respective populations being called significant [79]. While this difference could be originating from a true genetic effect specific to populations, more commonly it will greatly increasing the number of false positives due to spurious correlations. Various methods that estimate and control for the genetic relatedness structure between samples have been proposed in GWAS studies [80, 81, 82, 83, 84, 85]. The most commonly applied method is to include a select number of genotype principal components in the model [81], which has also been widely used when conducting eQTL analysis [20, 22].

1.2.3 Confounding Factor Effects

Consortium scale studies in which samples from multiple populations are processed across multiple laboratories are becoming the norm in eQTL datasets [18, 20, 22]. Additional to accounting for population structure, a particularly important aspect in analyzing these datsets is to correct for structures in the

samples arising due to non-genetic confounding factors that violate the independence assumption of the model. It has been shown that samples, which have been processed across different institutions or using different protocols, can be impacted by systematic variance that influence a large fraction of expressed genes. For example, factors such as non-specific binding due to probe design, and environmental differences such as atmospheric ozone levels [86] can be problematic in microarrays, while in RNA-seq GC content and insert size variations in the libraries can affect measurements [87]. If not correctly accounted for, these confounding factor effects can lower the power of the analysis by masking real genetic effects, and can also induce false positive results. Multiple groups have developed methods to estimate and correct for these confounding factor effects in eQTL analyses [66, 76, 77, 88, 89, 90, 91]. It has been shown that these increase the number of identified eQTL in practice. These methods can be broadly categorized into two groups.

Linear Fixed Effect Model The first group of methods attempt to model the confounding factor effects using certain assumptions about the factors and include the estimated covariates as fixed effects for correction [76,88,90,91]. In the basic linear fixed effect model, which is the most widely used method for mapping eQTL, the expression level of a single gene $\vec{y}$ for n individuals is modeled as linear combinations of genetic effects and covariates as

$$\vec{y} = \mathbf{X}\vec{\beta_x} + \mathbf{Z}\vec{\beta_z} + \vec{\epsilon} \tag{1.1}$$

where $\vec{y}$ is an $n \times 1$ vector of gene expression values, $\mathbf{X}$ is an $n \times g$ matrix of g genotypes and $\vec{\beta_x}$ is an $g \times 1$ vector of genotype coefficients. $\mathbf{Z}$ is an $n \times f$ matrix of f covariates and $\vec{\beta_z}$ is a $f \times 1$ vector of covariate coefficients. Here, confounding

factors are commonly estimated using the full expression matrix $\mathbf{Y}$, which spans the expression values $\vec{y}$ for every gene, and the estimated effects are modeled in $\mathbf{Z}$. Methods including principal component analysis (PCA) [88], factor analysis (FA) [76,91], and independent component analysis (ICA) [90] have been used in previous approaches to estimate confounding factors. Additional factors such as age, gender, and experimental batches, are also often included as covariates to account for their influence on expression level. Due to its closed form solution and simplicity, the linear fixed effect model has a very low computational cost and is relatively easy to interpret. However, as the number of parameters increase the statistical power of the model is reduced and there is also the potential of over correction.

Linear Mixed Model The second group of methods attempt to model a covariance structure between samples and include this information as a random effect in a linear mixed model [66,77,89]. The main advantage of a linear mixed model is that it can model dependent structures between samples, which can be problematic under the assumption of independence. Additionally, linear mixed models do not suffer from the loss of power due to additional parameters. In a linear mixed model framework, a random effect that accounts for correlation structures in the samples is used in addition to the fixed effects.

$$\vec{y} = \mathbf{X}\vec{\beta_x} + \vec{c} + \vec{\epsilon} \tag{1.2}$$

$$\vec{c} \sim N(\vec{0}, \tau^2\mathbf{K}) \tag{1.3}$$

$$\vec{\epsilon} \sim N(\vec{0}, \sigma^2\mathbf{I}) \tag{1.4}$$

Here, $\vec{y}$ is an $n \times 1$ vector of gene expression values, $\mathbf{X}$ is an $n \times g$ matrix of g genotypes and $\vec{\beta_x}$ is an $g \times 1$ vector of genotype coefficients. $\vec{c}$ is an $n \times 1$ random effect vector which is assumed to be sampled from a multivariate normal distribution with mean $\vec{0}$, and covariance $\tau^2\mathbf{K}$. $\vec{\epsilon}$ represents the random noise drawn from a multivariate normal distribution with mean $\vec{0}$ and covariance $\sigma^2\mathbf{I}$.

The key difference between confounding factor methods using this approach is the estimation of the $n \times n$ sample covariance matrix $\mathbf{K}$. Strategies from simply taking the covariance matrix of $\mathbf{Y}$ [66], to estimating $\mathbf{K}$ in a maximum likelihood framework [77, 89] have been used in previous studies. One particular concern in the application of linear mixed models in eQTL is that unlike the fixed effect model the linear mixed model does not have a closed form solution making the estimation of parameters more difficult. To address this issue, several approaches that increase the computational efficiency of parameter estimation have been proposed [82, 92].

1.3 Overview of Dissertation

In this thesis, we focus on analyzing gene expression profiles to estimate expression patterns that can be utilized to increase the accuracy of eQTL discovery in human datasets. Specifically, we use independent component analysis (ICA) to identify patterns in gene expression data and apply our findings to develop a linear mixed model confounding factor correction method. First, we introduce a new software tool that facilitates the application of ICA to gene expression data in chapter 2. In chapter 3 we describe CONFETI, a novel confounding factor

correction method we have developed, that specifically addresses the problem of estimating broad impact genetic effects. We then apply CONFETI to simulated data where we know the generative truth to evaluate the performance of CONFETI in comparison to other confounding factor correction methods in chapter 4. Finally, in chapter 5 we apply CONFETI and other confounding factor correction methods to a collection of human eQTL datasets and investigate the replication of eQTL as a measure of performance.

1.3.1 Analyzing Gene Expression Data with Independent Component Analysis

In chapter 2, we review the concept and applications of the widely used blind source separation method ICA. We then describe its main strength in analyzing gene expression profiles in comparison to PCA. Finally, we introduce `picaplot`, an R package that we developed for analyzing gene expression data with ICA. We describe the key features of `picaplot` including single and multi-run IC estimation, unsupervised detection of sample clusters, covariate association testing, and comprehensive visualization of results. Results obtained by analyzing publicly available gene expression data are presented to showcase the usage of `picaplot`.

1.3.2 CONFETI: An Independent Component Analysis Confounding Factor Correction Framework

In chapter 3, we introduce CONFETI, a novel linear mixed model confounding factor correction method based on ICA. CONFETI is a method designed to avoid the inclusion of genetic effects in the correction of confounding variation, thus maximizing the potential of broad impact eQTL discovery. This is implemented by estimating and filtering out candidate genetic effects from gene expression data using ICA. The non-genetic expression matrix is then used to construct a sample covariance matrix, which is used as a random effect in a linear mixed model framework. We describe the eQTL model in detail and provide a comparison to other published confounding factor correction methods.

1.3.3 Comparison of Confounding Factor Correction Methods Using Simulated Data

To first evaluate the performance of CONFETI and compare it to other published confounding factor correction methods, we use simulated eQTL data and analyze the results. We use yeast genotype data to simulate phenotype values for *cis*, *trans* and broad impact eQTL and use CONFETI and other published methods including PEER, ICE, and PANAMA to analyze the datasets. The results demonstrate that CONFETI most accurately estimates simulated eQTLs in the presence of broad impact eQTL and confounding factors. We present the results showing the overall accuracy of simulated eQTL and the recovery rate

in each simulated eQTL category.

1.3.4 Evaluating the Performance of Confounding Factor Correction Methods through Replicating eQTL in Human Data

In the final chapter, we present the analysis results of human eQTL datasets using CONFETI and other confounding factor correction methods. Since we do not have a gold standard eQTL dataset where a set of true eQTL are known to us, we focused on assessing the replication of *cis* and *trans*-eQTL in the analysis. For this purpose we used datasets obtained from the Multiple Tissue Human Expression Resource (MuTHER) consortium, which consisted of matched twin pairs for three different tissue types (Adipose, LCL, and Skin), and datasets obtained from the Genotype-Tissue Expression (GTEx) consortium, from which we selected 4 tissue pairs (Adipose - Subcutaneous and Visceral, Artery - Aorta and Tibial, Heart - Atrial Appendage and Left Ventricle, Skin - Suprapubic and Leg).

Accounting for confounding variation led to a significant increase of eQTL discoveries compared to simple linear regression, with linear mixed model based methods identifying the largest number of eQTL. However, we found little difference in identifying eQTL between linear mixed model confounding factor correction methods that accounted for the majority of the total variance in constructing the sample covariance matrix. Additionally, while a large frac-

tion of identified *cis*-eQTL replicated between twin and tissue pairs and across all datasets, most of the identified *trans*-eQTL were dataset specific and did not replicate well.

CHAPTER 2

ANALYZING GENE EXPRESSION DATA WITH INDEPENDENT COMPONENT ANALYSIS

2.1 Introduction

Genome-wide gene expression profiling by RNA-Seq or microarray is a prolific strategy for discovering novel biological pathways impacted by experimental treatments [93] and for revealing the impact of genetic variation on gene expression [41]. With the introduction of microarrays the simultaneous quantification of expression levels of many genes became easy, and RNA-seq expanded the scope by enabling the detection of previously unknown genes and isoforms. The number of measured genes varies depending on the measurement platforms but generally exceeds 20,000, creating multiple challenges in the analysis process related to the high-dimensionality.

Critical to drawing correct biological conclusions from the analysis of such data is pre-analysis detection and correction for systematic differences caused by measured and unmeasured factors, such as technical batch effects and heterogeneous environmental conditions [94]. If not accounted for, these cryptic factors can often contribute a significant proportion of the total variation in gene expression data, which can result in obscured signals of experimental impacts or worse, systematic biases and artifacts that are incorrectly interpreted as biological findings. Besides, these factors can potentially lead to interesting biological findings themselves by revealing differences in pathway activation patterns or

cellular states depending on different environments.

The most routinely used method to inspect the data for apparent patterns prior to analysis is principal component analysis (PCA). In most cases, the lower dimensional projections of the data onto the first few principal components are used to reveal patterns or clusters among samples that explain the largest amount of variance. While clearly a valuable approach due to the well defined statistical properties and relatively simple calculations, PCA is likely to return mixtures of multiple independent factor effects unless these happen to be aligned with the dimensions of greatest variation in the data [95]. Therefore, directly interpreting principal components can often be problematic and misleading. Additionally, by only considering an arbitrary number of components, patterns that have lower contribution to the overall variance can easily be missed.

Independent Component Analysis (ICA), which is a blind source separation method that decomposes the data into statistically independent components, can provide a clearer separation of these components. By using a stronger principle of statistical independence, ICA is expected to return interpretable components each aligning with an independent factor as long as these factors are non-Gaussian. ICA has been applied to problems such as voice and image separation, and more recently to high dimensional gene expression data to estimate non-Gaussian generative sources from an observed mixture. For example, studies have used ICA on problems such as expression pattern analysis [96, 97, 98, 99] , tumor classification [95, 100], and analyzing the effects of genetic variation [101, 102]. Here we propose a gene expression analysis frame-

work centered on ICA with information about every component to provide a more accurate, interpretable, and comprehensive estimation of covariate effects.

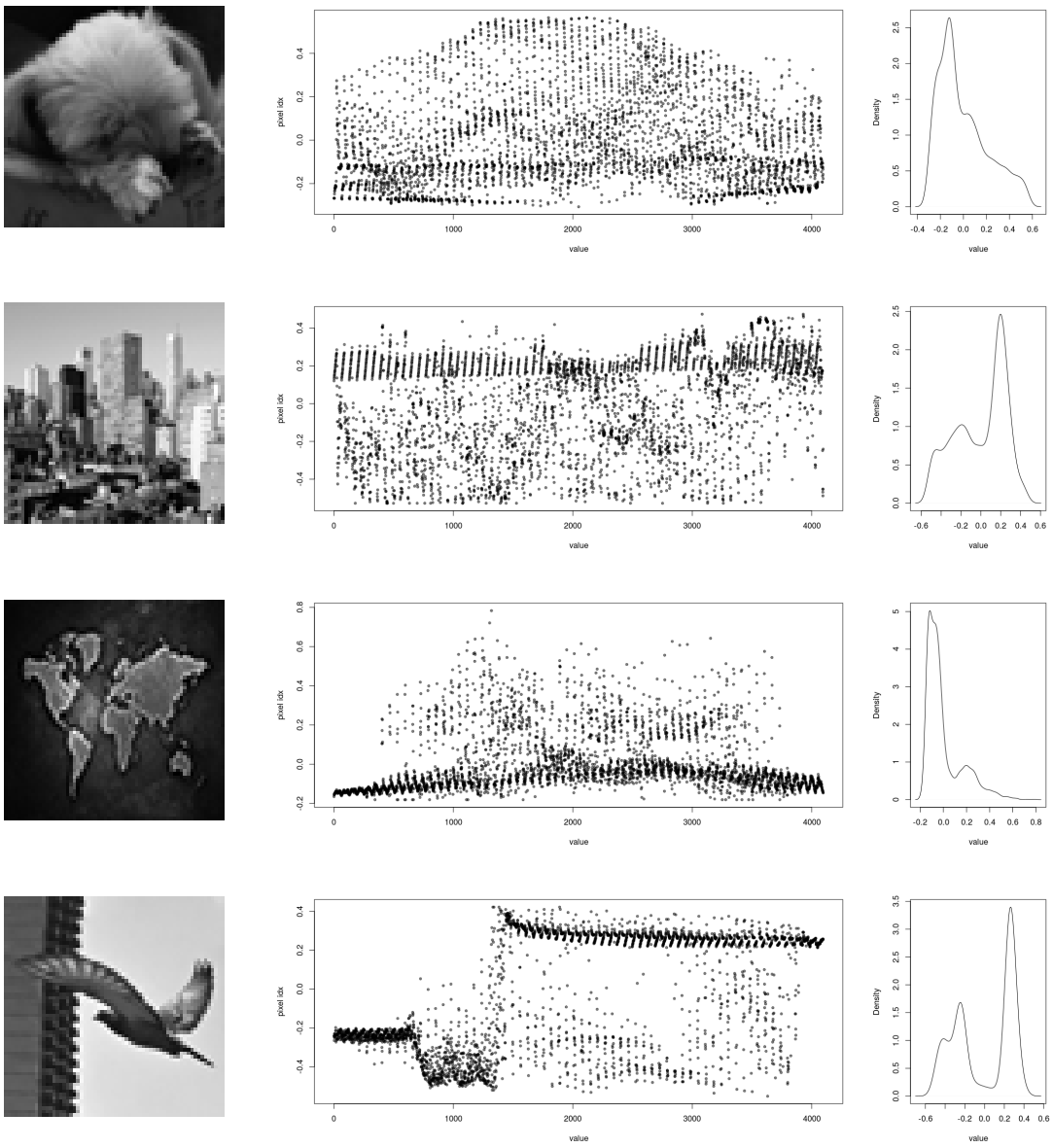
2.2 The Concept of Independent Component Analysis

Figure 2.1: Composite images

Examples of images created by mixing 4 different source images.

To explore the principles of ICA and how it is applied to gene expression data, let us first begin with a simple example using image separation. In Figure 2.1 a collection of images is shown made up of composite images generated by taking a weighted sum of 4 different 64 x 64 pixel gray scale source images. The source images used for this example are shown in Figure 2.2(a). Here, the problem that we are trying to solve with ICA is to estimated the generative sources from the composite images without any given information about the source images or the mixture weights. This framework is often referred to as blind source separation (BSS),since we are attempting to separate sources from a mixture without prior knowledge.

To explore this in the ICA framework we first need to find a mathematical representation of our problem. Since each image can be considered as a 4096



(a) (b) (c)

Figure 2.2: Four gray scale source images

(a) Images of our family dog, the view from my apartment, a world map, and a new york city pigeon. (b) Pixel brightness values for each image shown in a scatter plot. (c) Histogram of pixel brightness values for each image.

dimensional vector of pixel brightness values, as shown in Figure 2.2(b), we can

express each composite image as a linear equation:

$$\vec{m_i} = a_{i1}\vec{p_1} + a_{i2}\vec{p_2} + a_{i3}\vec{p_3} + a_{i4}\vec{p_4} \tag{2.1}$$

where $\vec{m_i}$ and $\vec{p_k}$, with $i, k \in \{1, 2, 3, 4\}$, are both 4096 dimensional vectors with the former representing composite images and the latter representing source images. Values of a_{ik} are showing the mixture weights of each image k in each mixture i.

Given that the mixture coefficients are a unique combination for each m_i without being a scalar multiple of another, one way to estimate each $\vec{p_k}$ would be to find a transformation of the observed $\vec{m_i}$.

$$\vec{p_k} = w_{k1}\vec{m_1} + w_{k2}\vec{m_2} + w_{k3}\vec{m_3} + w_{k4}\vec{m_4} \tag{2.2}$$

For this strategy to work, an assumption that defines the desired properties of the resulting $\vec{p_k}$ components is required. Without such a criteria to optimize for, any transformation is going to have equal importance making it impossible to converge on estimations for $\vec{p_k}$.

In the ICA framework, the two assumptions applied to accomplish this is that each $\vec{p_k}$ are statistically independent of each other and have non-gaussian distributions. This is based on the properties of the central limit theorem which states that the sum of independent and identically distributed random variables will tend to be more gaussian than their respective distributions. Here, the distributions of the source signals are not necessarily identical, but it has been shown in variations of the central limit theorem and in practice that this property still holds even with different distributions [103]. Thus, estimating the most non-Gaussian components from the observed mixture, we would be

able to recover the components closest to the source signals. Based on this criterion, ICA estimates the values for w_{kj} that maximizes the non-gaussianity of the resulting $\vec{p_k}$. To be more specific, ICA will look for an orthogonal projection matrix $\mathbf{W}$, constructed from values of w_{kj}, that maximizes the non-gaussianity of the resulting pixel brightness distributions. We can see in Figure 2.2(c) that the distribution of pixel brightness values are non-Gaussian for the source images, which enables the application of ICA to this problem. The coefficient values a_{ik} can then be calculated by taking the inverse of $\mathbf{W}$. Figure 2.3 shows the estimated source images by ICA in comparison to the PCA results. We can see that the results obtained by ICA show a clearer separation of the source images, while the principal components show little difference from the observed mixtures.

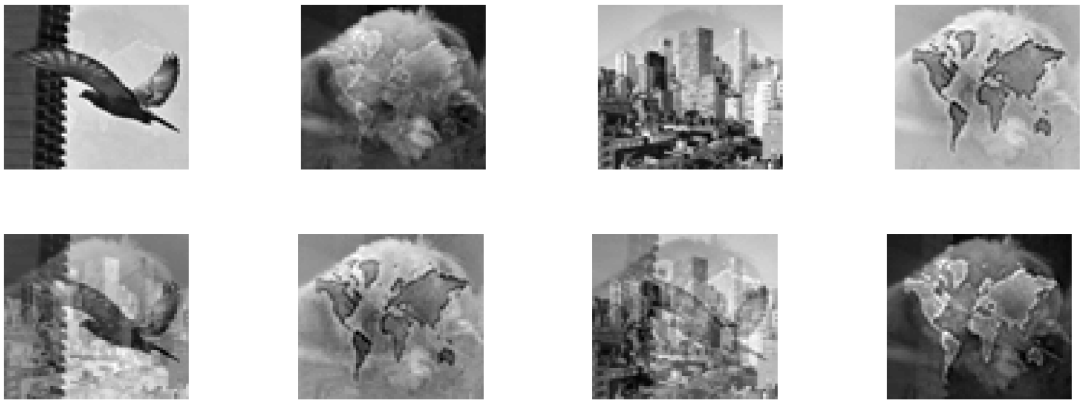


Figure 2.3: Estimated source images by ICA and PCA

The resulting components obtained by applying ICA (top row) and PCA (bottom row) to the image mixtures.

Similar to the image mixture example, we can project the idea directly to gene expression profiles. We hypothesize that each expression profile of a single sample i is a combination of multiple independent components, which can be

either genetic or non-genetic components.

$$\vec{y_i} = a_{i1}\vec{s_1} + a_{i2}\vec{s_2} + \cdots + a_{ik}\vec{s_k} = \sum_{j=1}^{k} a_{ij}\vec{s_j} \tag{2.3}$$

where $\vec{y_i}$ is a vector of g gene expression values for a single sample, and $\vec{s_j}$ are g-dimensional independent components of gene weights that are shared among all samples and the scalar component coefficients a_{ij} represent the contribution of each independent component $\vec{s_j}$ for sample i. In this case, the estimation of independent components will depend on the assumption that gene weight distributions of each $\vec{s_j}$ are non-gaussian. This is not an unreasonable assumption, since most biological processes will be strongly driven by a subset of the total genes resulting in super gaussian distribution, and non-genetic effects such as batch effects have a tendency to have a broad impact thus resulting in a heavy tailed distribution.

There are several algorithms that use different approaches to perform ICA, such as maximum likelihood estimation [104], high-order correlations optimized by Jacobi algorithms [105], reproducing kernel Hilbert space based approach [106], and estimation based on efficient entropy estimation [107]. In this thesis, we used the well studied FastICA algorithm [108] for efficient and robust computations, which uses an approximation of negentropy as a measure of statistical independence.

Here we briefly review the estimation of negentropy used as an approximation of non-gaussianity which is used in the FastICA algorithm [108]. Negentropy is a measure of the departure of a given distribution from a gaussian random variable with the same mean and variance (or covariance structure). It is based on differential entropy H, which for a random variable x with density

p_x can be shown as

$$H(x) = -\int p_x(\eta) \log p_x(\eta) d\eta \tag{2.4}$$

This can be considered as a measure of average surprisal of a random variable, or a measure of how structured a random variable is. It takes a small value when a variable is more predictable and structured, for example if the probability of the majority lies close to 0 or 1. It has been shown in information theory that for random variables of equal variance, a gaussian variable is the least structured with the largest entropy [103]. Based on this result, we can use the difference between the differential entropy of a given random variable and a gaussian variable with equal variance as a measure of non-gaussianity. This measure is called negentropy and is defined as

$$J(x) = H(g) - H(x) \tag{2.5}$$

where g is a gaussian random variable with $var(x)$. Since the gaussian variable will always have the maximum differential entropy, negentropy will always be positive for a non-gaussian x, and 0 if x is gaussian. Despite having a statistical justification negentropy is difficult to calculate computationally since it needs an estimate of the probability density function. Thus, an approximation of negentropy based on higher order statistics is used. Hyvarinen proposed an approach that uses a generalized form of higher-order cumulant approximation, in which negentropy is approximated by

$$J(x) \propto [E\{G(x)\} - E\{G(g)\}]^2 \tag{2.6}$$

where the following nonquadratic functions G with $1 \leq a_1 \leq 2$ are suitable choices [109].

$$G_1(x) = \frac{1}{a_1} \log \cosh a_1 x \tag{2.7}$$

$$G_2(x) = -\exp(-x^2/2) \tag{2.8}$$

After a nonquadratic function is selected, the FastICA algorithm finds a projection matrix that maximizes the negentropy using a fixed-point iteration scheme. While it is possible to estimate components one at a time, this could lead to cumulated errors in the components estimated later in sequence. Therefore, we chose the simultaneous estimation of components, which estimates a given number of components in parallel.

2.3 `picaplot`: an R package for Identifying Cryptic Covariates in Genome-Wide Gene Expression Data

For the specific purpose of applying ICA to gene expression data, we developed a publicly available R package `picaplot`. The key features of `picaplot` include simple application of ICA to gene expression data, automated cluster detection to identify cryptic covariate effects and interpretable outputs that are easy to incorporate as fixed effects in analyses using linear models. The package also implements parallel functionality for PCA to compare the results with that of the ICA output.

2.3.1 Single-run and Ensemble ICA Estimation

The observed expression values $\mathbf{Y}$ are assumed to be linear combinations of non-Gaussian statistically independent components and are decomposed into a

mixing matrix **A** and component matrix **S**:

$$\mathbf{Y} = \mathbf{AS} \tag{2.9}$$

where **Y** is an $n \times g$ matrix with expression values for n individuals and g genes. **A** is an $n \times k$ mixing matrix with k component coefficients for each sample, and **S** is a $k \times g$ independent component matrix with g gene weights for each component.

The number of components k has to be set prior to decomposition and `picaplot` provides two ways to estimate k. The first approach is based on the % variance explained by principal components. The user can either directly set k or can provide the amount of variance to be included and `picaplot` will automatically determine the number of components based on the % variance. When neither k nor the % variance are provided by the user, the `run_ica()` function will automatically use the number of principal components that explain a 99% of the total variance by default.

In the second approach, `run_ica()` is executed multiple times and the similarity between components is estimated based on the gene weights of **S** in each run. This is to address the problem of unstable results obtained from a single estimation of ICA. Since the optimization process begins with a random initialization of the projection matrix, there is a possibility that local maximums could lead to different results in each run. One way to estimate the similarity between components is to calculate the pair-wise correlation of each component. However, given the high-dimensionality of the components this could lead to a systemic underestimation of correlation values in cases where there are only few genes that have a significant gene weight. Therefore, the recommended way of calculating the similarity between components is to only use the gene weight

values of ”peaks”, which are gene weights that are greater than 2 standard deviations of the corresponding component gene weight distribution. Based on the calculated similarity the estimated components are grouped together by hierarchical clustering, and clusters with a total number of components exceeding 90% of the number of runs are selected. For example, if ICA was performed 10 times, clusters with more than 9 components are considered as replicating. This arbitrary threshold can be set by the user, however we recommend a stringent threshold for robust estimation of replicating components. The average of component gene weights within each replicating cluster are then used to generate an ensemble estimate of replicating components.

2.3.2 Covariate Association Checking

The `covariate_association_check()` function can be used on the ICA decomposition results to identify associations between the IC coefficients (columns of **A**) and measured covariates. Since IC coefficients show the relative contribution of each IC in each sample, this information can link estimated ICs to measured variables of interest. For example, this could estimate the effects of commonly included covariates such as age, gender and experimental batches. Moreover, if covariates for experimental treatments are available the association analysis could reveal ICs related to specific biological pathways.

2.3.3 Cluster Detection in IC coefficients

For cases where covariate measurements are limited or not present, we have included a feature for model-based clustering to detect distinct clusters in the IC coefficients. This could potentially reveal interesting structures that separate samples into different groups without the need of any measured labels or traits. This can be particularly useful in cases where unmeasured differences in sample processing steps or unknown environmental differences have systematic influences on the expression measurements of genes. Implemented in the `detect_clusters()` function, the feature is based on functionalities of the `mclust` R package [110]. It uses an Expectation-Maximization (EM) algorithm to perform a maximum-likelihood estimation of parameterized Gaussian mixture models on the univariate IC coefficient data for each component, and selects the ideal model based on the Bayesian Information Criterion (BIC) [111]. In other words, it fits different numbers of Gaussian distributions using the IC coefficient estimates of each component and finds the number of Gaussian distributions that best explain the data. The function returns the estimated number of clusters for each IC coefficient with their corresponding cluster labels if multiple clusters are estimated.

2.3.4 Correcting IC Effects in Linear Models

To include the IC coefficients associated with known covariates or those with multiple clusters in a linear model as fixed effects, a matrix of IC coefficients can be generated by the `get_covariate_mx()` function. This will automatically

select the IC coefficient values which show a significant association with a tested covariate or estimated to have multiple clusters through `detect_clusters()`. A custom matrix of selected ICs can also be easily retrieved in case the user wishes to correct for IC effects that are not associated with covariates and also not estimated to have multiple clusters. By incorporating these coefficients in a linear model, the effects of factors that are not directly related to the variable being tested can be corrected out to increase the statistical power of the analysis.

2.3.5 Visualization of Results

The results generated by ICA decomposition represent a high dimensional structure, especially estimated components in **S**, thus effective visualization is crucial in inspecting and analyzing the results. In `picaplot`, each IC is visualized by combining the positional information of genes with their gene weights, accompanied by the corresponding IC coefficients. To combine these results for multiple components, we have implemented a report generating function `report_gen()`, which generates an `HTML` report containing gene weight plots and IC coefficient plots with information regarding associated covariates and key driver genes for every component.

2.3.6 Application

To demonstrate the features of `picaplot` we analyzed a publicly available gene expression dataset described in [112]. We downloaded the dataset from the Gene Expression Omnibus (GEO), using the accession number GSE60028. After downloading the expression data and covariates, we first filtered the expression values to genes that are mapped to chromosomes 1 to 22, X and Y which left us with 26419 genes measured for 47 samples. From the recorded covariates, we selected gender, tested allergen, and the level of reaction to test the association with estimated ICs. The example dataset that has been used is included in the `picaplot` package and can be accessed by `data(expr_data, sample_info, probe_info)`.

2.4 Results

42 independent components were estimated, based on the observation that 42 principal components explained more than 99 % of the total variance. We found that the component estimated by ICA associated with the gender status of the samples had the most significant covariate association p-value and it separated the samples almost perfectly into two clusters, while the principal component with the strongest association to gender showed a weaker separation (Figure 2.4). The gene weight plots show that ICA identified fewer genes that highly contribute to the component in comparison to PCA (357 versus 1198), which could be explained a better separation of composite effects by ICA.

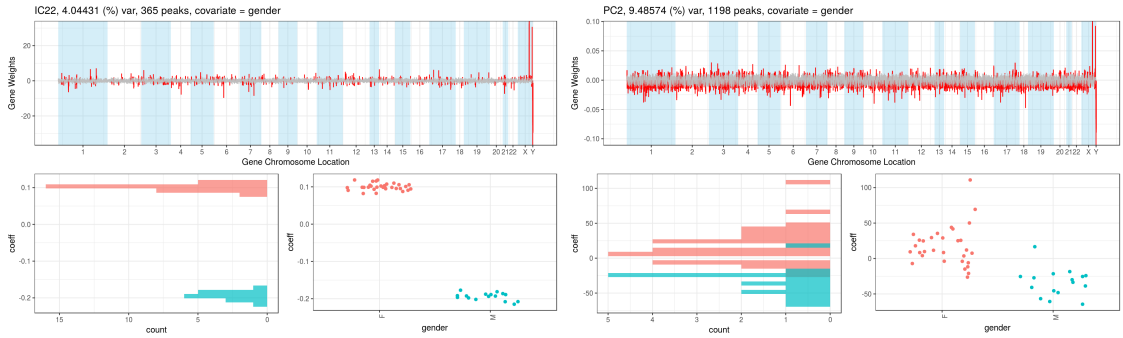


Figure 2.4: Gender specific expression patterns estimated by ICA and PCA

`picaplot` visualizations of gene weight loadings showing highly contributing genes in red and low contributors in grey ordered by their position in the genome (top) and coefficient/projection visualization plots showing sample clustering colored by the associated covariate (bottom). Components estimated by ICA (left) and PCA (right) showing the components with the most significant association with the known covariate gender.

We were able to replicate this finding by identifying components that were highly associated with the gender status in a microarray dataset of smokers and non-smokers, and an RNA-seq dataset obtained from the GTEx consortium. The separation was clearer in microarray compared to RNA-seq, but in both cases the separation was better than components estimated by PCA (Figure A.1). This demonstrates that ICA can robustly estimate the gender specific effects from gene expression profiles. These components could be used to estimate the gender status if the covariate is missing, or as a quality control check to detect any mis-labeled samples if the gender status was recorded.

We were also able to identify multiple components associated with allergens, with the most significant associated component strongly separating the samples between petrolatum and nickel as reported in the study. Additionally, ICA returned multiple components showing no association with known covariates but estimated to have multiple clusters (Figure 2.5).

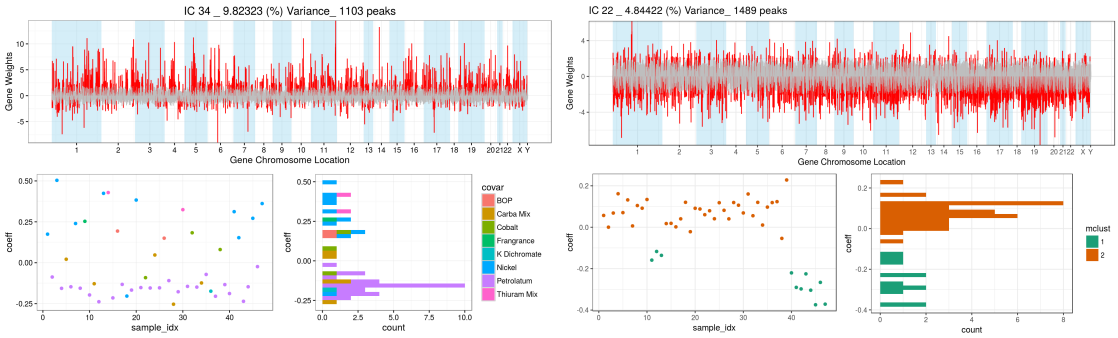


Figure 2.5: Allergen associated component and example of cluster estimation

Estimated IC associated with the allergen covariate is shown on the left. An example of a component with multiple clusters detected by `detect_cluster()` with no known covariate association is shown on the right.

2.5 Conclusion

Identifying and accounting for cryptic covariates when analyzing genome-wide gene expression data is critically important for correct biological discovery. Given that the cryptic covariates discovered by ICA are often expected to be distinct from those returned by PCA, we highly recommend applying both methodologies to any experiment. Not only can this information be used to correctly account for hidden sources of variation, it can also be used to improve the design of future experiments by understanding potential factors influencing gene expression levels. In summary, `picaplot` provides an easy-to-use, yet comprehensive workflow to use both methods to inspect genome-wide gene expression data for cryptic covariates and to correct for and model these covariates in subsequent analyses.

CHAPTER 3

CONFETI: AN INDEPENDENT COMPONENT ANALYSIS CONFOUNDING FACTOR CORRECTION FRAMEWORK

3.1 Introduction

Systematic variation introduced by non-genetic factors, such as technical variation caused by differences in laboratory procedures or distinct study environments are major challenges in analyzing high dimensional gene expression measurements [8, 94, 113]. These can be especially problematic in detecting small genetic effects in eQTL analysis, by obscuring true effects and lowering the statistical power. To address this issue, multiple studies have developed confounding factor correction methods based on various strategies [66,76,77,88,89,91,114,115,116]. Generally, these confounding factor methods account for non-genetic variation in eQTL studies by learning and modeling systematic variation directly from the multivariate structure observed in gene expression data. When used in combination with corrections for population structure [89], confounding factor analysis can both increase power in eQTL studies and reduce false positives by accounting for non-genetic factors that impact many genes. While confounding factor analyses should increase the correct discovery of both *cis*- and *trans*-eQTL by increasing detection power [17, 113], a known problem of all confounding factor methods is the potential to model the effects of broad impact eQTL as confounding variation [77, 117]. Previous approaches to avoid the removal of broad impact eQTL as confounding factors include, jointly estimating the error structure with genetic information [77],

and using only a subset of genes to estimate the confounding structure [115]. However, such approaches do not explicitly identify individual confounding factors and could generate different results based on selected genes, which is a non-optimal strategy for avoiding the removal of variation produced by broad impact eQTL.

In this chapter, we describe a new framework that is designed to improve on the performance of confounding factor methods to identify broad impact eQTL. The CONFETI (CONfounding Factor Estimation Through Independent component analysis) framework makes use of ICA, described in the previous chapter, to separate genetic components from non-genetic components learned from multivariate gene expression variation. CONFETI takes advantages of the key strength of ICA to estimate generative sources of variation from an observed mixture, which can be used to separate independent sources of variation, such as genetic versus non-genetic factors. After these generative sources have been estimated by ICA, CONFETI automatically filters out those that are candidates for broad impact eQTL variation and retains the rest as a lower dimensional representation of the non-genetic confounding variation. By explicitly identifying clear candidate signals of broad impact eQTL, CONFETI prevents the modeling away of true genetic effects and increases the discovery potential of confounding factor analyses.

3.2 Design of the CONFETI Framework

The CONFETI framework is constructed to systematically avoid the tendency of other confounding factor analysis methods to model broad impact eQTL as confounding variation. This is accomplished by leveraging ICA to identify generative sources of multivariate gene expression variation and then screening candidates based on component correlations with genotypes, which are then omitted from the confounding factor correction (Fig 3.1). The reason ICA is particularly well suited for identifying candidate broad impact eQTL is that the method is designed to separate independent sources of multivariate variation.

ICA assumes that the observed data for each sample is a linear combination of non-Gaussian statistically independent components. When applying ICA, the vector of expression values for an individual are modeled as weighted sum of independent components:

$$\vec{y_i} = a_{i1}\vec{s_1} + a_{i2}\vec{s_2} + \cdots + a_{ik}\vec{s_k} = \sum_{j=1}^{k} a_{ij}s_j \tag{3.1}$$

where $\vec{y_i}$ is a g-dimensional vector of gene expression values for a single sample, and independent components $\vec{s_j}$ are g-dimensional vectors of gene weights that are shared among all samples and the scalar component coefficients a_{ij} represent the contribution of each independent component $\vec{s_j}$ for sample i (Fig 3.1). When considering all samples together, the above can be simply expressed as a matrix decomposition:

$$\mathbf{Y} = \mathbf{AS} \tag{3.2}$$

where $\mathbf{Y}$ is an $n \times g$ matrix with i^{th} row $\vec{y_i}$. $\mathbf{A}$ is the $n \times k$ mixing matrix with the j^{th} column holding component coefficients $\vec{a_j}$ for component j, and $\mathbf{S}$ is the

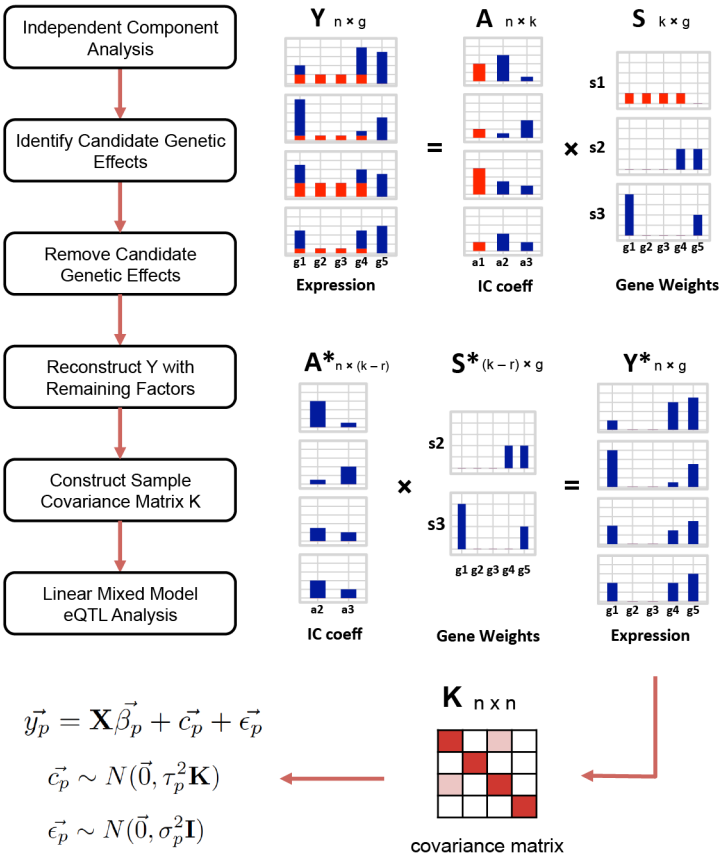


Figure 3.1: The CONFETI framework

ICA is used to decompose the gene expression matrix **Y** into an IC coefficient matrix **A** and a component matrix **S**. Associations between the genotypes and coefficients in matrix **A** are tested to label any candidate genetic effects to be removed from the correction. In the example above, the first IC, shown in red, is marked as a candidate genetic component and the corresponding columns of **A** and rows of **S** are removed. Using the lower rank $\mathbf{A}^*$ and $\mathbf{S}^*$, expression values originating from non-genetic components are reconstructed in $\mathbf{Y}^*$. Finally, **K** is created by calculating the sample covariance matrix of $\mathbf{Y}^*$, and included as a random effect in the mixed model for eQTL analysis.

$k \times g$ independent component matrix in which the j^{th} row is $\vec{s_j}$. **A** and **S** are estimated by finding a projection of **Y** that maximizes the non-gaussianity of the gene weight distribution of each row in **S**. In CONFETI these are identified by using the FastICA algorithm for reliable and fast computation [108].

Since ICA recovers factors by assessing non-gaussianity and not the amount of variation explained as in methods such as Principal Component Analysis (PCA) or any other factor analysis method [103], ICA is able to more clearly resolve separate factors responsible for variation, while a PCA or factor analysis will tend to identify composite effects, which are likely to be mixtures of multiple factors. The critical assumption for application of ICA in the CONFETI framework is that broad impact eQTL will have non-Gaussian impacts on the multivariate expression profile and that the effects of these eQTL will be relatively independent of other genetic and non-genetic factors. Complete independence is not necessary, since the framework only has to identify and retain enough of the expression variation due to a broad impact eQTL to make it detectable with an association test. The assumption that broad impact eQTL will tend to have non-Gaussian impacts is not particularly restrictive given that we expect eQTL with large enough effects to impact only a subset of the total number of genes and therefore be detectably non-Gaussian. The assumption that broad impact eQTL are relatively independent of each other is also not overly restrictive in humans given the low linkage disequilibrium observed among non-local genotypes throughout the genome. While the assumption that broad impact eQTL are largely independent of non-genetic factors is not always expected to hold, it seems likely in many cases unless there is a reason to expect broad impact eQTL to strongly interact with non-genetic factors such as sample-specific environmental effects or technical effects such as differences between laboratories and procedures. Furthermore, in cases where broad impact eQTL are completely conflated with non-genetic factors, these broad impact eQTL will be indistinguishable from non-genetic contributions to the observed multivariate gene expression variation and will be modeled away by any confounding

factor method. In sum, the only accurately detectable broad impact eQTL are those that have properties that are expected to make them identifiable by ICA.

The complete CONFETI framework involves running ICA on the multivariate gene expression data, an automated detection step to identify candidate broad impact eQTL by assessing associations with genotypes, and omission of these factors for the construction of the random effect sample covariance matrix used in a mixed model confounding factor analysis (Fig 3.1). While this approach could be used in combination with confounding factor methods that use a fixed covariate approach [88,90,91,116,118,119,120], the framework more naturally integrates with mixed model approaches to confounding factor analysis, which do not suffer from the loss of power due to increased number of parameters in the model. A covariance matrix constructed from the non-genetic independent components is used to model confounding factors as random effects in a linear mixed model eQTL approach.

We note that our framework differs from ICA methods for eQTL detection that treat the identified ICs as meta-genes, where these methods cannot reliably distinguish the specific gene effects of individual eQTL [101,102]. The only method that we are aware of close to this framework is ISVA, which uses ICA within the Surrogate Variable Analysis (SVA) method for iteratively modeling pre-specified fixed effects and confounding variation [90]. ISVA is not appropriate for eQTL analysis since it begins the iterative approach by pre-specifying the fixed effects and therefore pre-supposing the existence of a relationship, which would introduce a bias towards finding eQTL false positives. CONFETI on the other hand uses ICA to separate candidate broad impact eQTL without the

need of pre-specifying the existence of the eQTL. We also note that in the mixed model based method PANAMA [77], the authors discuss a strategy for avoiding the over-correction of *trans*-eQTL by jointly estimating the covariance matrix with genotype effects to avoid including those effects in the correction [77]. However, this approach is not a feature of PANAMA included in the LIMIX package [121], which the authors have directed us to use. Moreover, the information that can be obtained from the gene loadings are not considered in the estimation step of PANAMA by being integrated out. In summary, the CONFETI framework utilizes the optimal properties of ICA to detect broad impact eQTL by excluding genetic effects from confounding variation accounted for in a mixed model, thereby taking advantage of the performance increases provided by mixed model confounding factor analysis without reducing the ability to identify broad impact eQTL.

3.3 Methods

3.3.1 Independent Component Analysis

To apply ICA to gene expression data and generate a sample covariance matrix, we developed a custom R package `confeti` which is publicly available at `https://github.com/jinhyunju/confeti`. The independent component estimation features are using functions adopted from the fastICA R package [122] which implemented the computationally efficient and robust FastICA algorithm [108] based on a fixed-point algorithm to find directions maximiz-

ing the negentropy to identify statistically independent components (ICs). The number of ICs that can be estimated is the smaller of the sample size or the number of features (genes), and the sign of any particular estimated component is arbitrary. As the estimated ICs do not have any particular order and have the potential to change based on the input of number of components to estimate [90, 100, 123], the package supports diagnostics for assessing optimal IC number such as functionality to estimate replicating ICs between multiple runs for ensemble ICA estimation. In our analyses, we used the number of principal components that accounted for 95% of the data variance as the number of components to be estimated to provide a fair comparison to PANAMA [77].

3.3.2 Removal of Candidate Broad Impact eQTL

After decomposing the observed data $\mathbf{Y}$ into $\mathbf{A}$ and $\mathbf{S}$ we test for any significant associations between the component coefficients (columns of $\mathbf{A}$) and all genotypes. As in fixed effect eQTL models, we fit a linear regression model with the IC coefficient as the dependent variable and the genotype values as independent variables. After calculating p-values for each IC coefficient and genotype pair we identified candidate broad impact eQTL using a Bonferroni corrected p-value threshold of 0.05.

After filtering out r ($0 \leq r < k$) components with significant genotype association, we reconstruct expression matrix $\mathbf{Y}^*$ originating from non-genetic factors using the remaining $k - r$ components:

$$\mathbf{Y}^* = \mathbf{A}^*\mathbf{S}^* \tag{3.3}$$

where $\mathbf{Y}^*$ is an $n \times g$ matrix, $\mathbf{A}^*$ is a $n \times (k-r)$ matrix and $\mathbf{S}^*$ is a $(k-r) \times g$ matrix.

3.3.3 Construction of the Sample Covariance Matrix

We used two approaches to construct the sample covariance matrix $\mathbf{K}$ for the random effect part of the mixed model. Our first approach was to use a simple location-scale normalization of each gene of $\mathbf{Y}^*$:

$$Z^*_{ip} = (Y^*_{ip} - \mu_p)/\sigma_p \tag{3.4}$$

and then calculate sample covariance matrix:

$$\mathbf{K} = \text{cov}(\mathbf{Z}^*) \tag{3.5}$$

We label this approach CONFETI-I since it can be thought of as a specific, lower dimensional approach to Intersample Correlation Emended (ICE), one of the first methods to estimate a sample structure for confounding factor analysis [66] by estimating the sample covariance matrix using the full dimensional observed expression data.

For our second approach, we couple CONFETI with PANAMA (Probabilistic ANAlysis of genoMic dAta) [77] that estimates the covariance structure using a maximum likelihood framework. Using this approach, the likelihood objective can be stated as:

$$p(\mathbf{Y}^*|\mathbf{K}_{\text{panama}}) = \prod_{p=1}^{g} \mathcal{N}(\vec{y^*_{.p}}|\tau^2_p\mathbf{K}_{\text{panama}} + \sigma^2_p\mathbf{I}) \tag{3.6}$$

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{C}}) = \text{argmax}_{\boldsymbol{\theta},\mathbf{C}}\ p(\mathbf{Y}^*|\mathbf{C}, \boldsymbol{\theta}) \tag{3.7}$$

where $\mathbf{C}$ is an $n \times Q$ matrix initialized by projecting the observed data onto the first Q principal components explaining 95% of the variance, and $\boldsymbol{\theta}$ is the set of hyperparameters consisting of $\{\{\alpha_q^2\}, \sigma_p^2\}$. Each α_q^2 then represents the weight of the q^{th} column of $\mathbf{C}$, $\mathbf{C}_{\cdot q}$ in constructing the sample covariance matrix:

$$\mathbf{K} = \sum_{q=1}^{Q} \hat{\alpha_q^2} \hat{\mathbf{C}}_{\cdot q} \hat{\mathbf{C}}_{\cdot q}^{\mathrm{T}} \tag{3.8}$$

We label this approach CONFETI-P, where we use of the implementation of PANAMA included in the LIMIX package [121] for the estimation of $\mathbf{K}$.

3.3.4 Linear Mixed Model eQTL Analysis

We model the genetic effects from SNPs and covariates as fixed effects and confounding factor effects as random effects, such that the expression levels for gene p in n individuals are:

$$\vec{y}_p = \mathbf{X}\vec{\beta}_p + \vec{c_p} + \vec{\epsilon_p} \tag{3.9}$$

$$\vec{c_p} \sim \mathcal{N}(\vec{0}, \tau_p^2 \mathbf{K}) \tag{3.10}$$

$$\vec{\epsilon_p} \sim \mathcal{N}(\vec{0}, \sigma_p^2 \mathbf{I}) \tag{3.11}$$

Where n is the number of samples, g the number of genes, s the number of SNPs, and v the number of covariates. Each gene expression vector $\vec{y}_p$ has dimension $n \times 1$ and is mean centered. The $n \times (s + v)$ genotype and covariate matrix $\mathbf{X}$ indicates the number of minor alleles for each SNP coded as 0,1,2 and any additional covariates. $\vec{\beta_p}$ is the $(s + v) \times 1$ dimensional coefficient vector representing the fixed effect of the SNPs and covariates on gene p. The confounding effect is included in the model as a $n \times 1$ random effect $\vec{c_p}$ sampled from a multivariate

normal distribution with covariance $\tau_p^2\mathbf{K}$, where $\mathbf{K}$ is the $n \times n$ sample covariance matrix constructed the corresponding confounding correction method, τ_p^2 is a scalar weight for $\mathbf{K}$ in the random effect, and $\vec{\epsilon_p}$ is a $n \times 1$ vector representing the independent error for gene p with scalar weight σ_p^2.

3.4 Conclusion

In this chapter, we have described our novel linear mixed model confounding factor correction method CONFETI. The focus of CONFETI is to correctly distinguish genetic variation from non-genetic variation to maximize the potential of broad impact eQTL discovery. In the following chapters, we compare the performance of CONFETI with other linear mixed model based confounding factor correction methods including PEER, ICE, and PANAMA. We first evaluate the performance in simulated data where the ground truth is known to us.

CHAPTER 4

COMPARISON OF CONFOUNDING FACTOR CORRECTION METHODS USING SIMULATED DATA

4.1 Introduction

While the number of reported eQTL findings are increasing in various tissue types across multiple human populations, only a very limited number of identified eQTL have been validated through experiments [124, 125]. Since a gold standard dataset for known eQTL is not available, we first evaluated the performance of multiple confounding factor correction methods using simulated data in which the generative truth is known to us. We simulated eQTL data based on yeast genotypes used in previous studies [126], in order to run multiple simulations with manageable data sizes. More specifically, we simulated *cis*, *trans*, and broad impact eQTL, where *cis* and *trans*-eQTL are single gene-genotype pairs, and broad impact eQTL are single genotypes associated with multiple genes. By simulating these three eQTL categories, we investigated the balance of each confounding factor correction method between correctly accounting for non-genetic variance and retaining true genetic signals.

We compared CONFETI-I and CONFETI-P, described in the previous chapter, to simple linear regression with no confounding factor correction (LINEAR), a widely used confounding factor method probabilistic estimation of expression residuals (PEER) [76], and mixed model confounding factor methods ICE [66] and PANAMA [77]. Additionally, we also considered a strategy based on CON-

FETI but substituting PCA for ICA, otherwise applying exactly the same approach to removal of principal components with significant genotype associations and calculating **K** for the remaining components weighted by the variance they explain, an approach we labeled PCAKMX. We evaluate the performance of each method by calculating the Area Under the Curve (AUC) for Receiver Operating Characteristics (ROC) curves, and by investigating the True Positive Rate (TPR) for varying False Discovery Rate (FDR) thresholds for the three categories of simulated eQTL.

4.2 Methods

4.2.1 eQTL Simulation

To mirror real cases where a reasonable number of broad impact eQTL have been repeatedly identified, we used yeast as a model [72, 73, 74]. 2956 yeast genotypes from the study of Smith et al. [126] and randomly sampled 3000 yeast gene annotations were used to create simulated datasets. The genome coordinates of the genotypes and sampled genes were used to simulate *cis* and *trans*-eQTL relationships. First, a matrix with a dimension of number of genotypes × number of expression phenotypes was created that marked genotype and phenotype pairs *cis* if the starting position of the gene and the genotype were within 100,000 bases and *trans* if the distance was greater. From this matrix we sampled 2500 genotype and phenotype pairs which consisted of 80% *cis* and 20% *trans* relationships. In total, for each simulated dataset, we included

2000 *cis*-eQTL, 500 *trans*-eQTL, and 10 broad impact eQTL. We simulated each broad impact eQTL to affect 10% of the expression phenotypes. Effect sizes for *cis*-eQTL were sampled from $\mathcal{N}(0.8, 1)$ and effect sizes for *trans*-eQTL and broad impact eQTL were sampled from $\mathcal{N}(0.48, 1)$ (70% attenuation of *trans*-effects) to reflect observed effect sizes in real data. After the eQTL effects were simulated, we added normally distributed random noise sampled from $\mathcal{N}(0, 1)$. For confounding factor effects, we simulated two types of confounding factors: sparse and dense. For sparse confounding factors 30% of phenotypes were affected with effect sizes drawn from $\mathcal{N}(1, 0.5)$, and for the dense confounding factors, the effect over all genes followed a standard normal distribution $\mathcal{N}(0, 1)$. We tested 2 scenarios, each with 30 confounding factors: sparse only, and mixed (15 sparse and 15 dense). We simulated and analyzed 50 datasets for each of these two scenarios, a total of 100 datasets.

4.2.2 Confounding Factor Correction Methods

We compared CONFETI-I and CONFETI-P, as described in the previous chapter, to other published confounding factor correction methods PEER, PANAMA, and ICE. Here, we briefly review the confounding factor correction methods.

PEER As a bayesian factor analysis model, PEER estimates Gaussian factors from a given expression matrix and uses automatic relevance detection (ARD) to discard factors of minimal importance [76].

$$P(Y_{ip}|\mathbf{f}_i, \mathbf{w}_p, \tau_p) = \mathcal{N}(Y_{ip}|\sum_{k=1}^{K} w_{p,k} f_{k,i}, \frac{1}{\tau_p}) \tag{4.1}$$

$$P(w_{p,k}|\beta_k) = \mathcal{N}(w_{p,k}|0, \frac{1}{\beta_k}) \tag{4.2}$$

$$P(f_{k,i}) = \mathcal{N}(f_{k,i}|0, 1) \tag{4.3}$$

As shown in the above equation, each gene expression level for individual i and gene p, $\mathbf{Y}_{ip}$, is modeled as a linear combination of k Gaussian factors $\mathbf{w}_p$ and their associated gaussian weights $\mathbf{f}_i$, with gamma priors on noise precisions τ_p and β_k. Estimated factor weights can be included in the model as fixed effects or a residual matrix can be calculated by subtracting the factor effects from the phenotype matrix. Here, we chose to include the estimated factor weights as covariates in the linear fixed effect model and used 25% of the sample size as the initial number of factors as recommended by the authors [120].

ICE In the ICE framework [66], the sample correlation structure is used to approximate the confounding factor effects. First, the expression values are mean centered and divided by the standard deviation, and the sample covariance matrix is calculated by the covariance of the normalized expression values. This is analogous to substituting the lower dimensional $\mathbf{Y}^*$ with the full expression value matrix $\mathbf{Y}$ in equations 3.4 and 3.5.

PANAMA A maximum likelihood approach is used to estimate factors that best explain the phenotype covariance structure in PANAMA [77]. Similar to ICE, considering the full expression matrix $\mathbf{Y}$ instead of $\mathbf{Y}^*$ in equations 3.6, 3.7, and 3.8 outlines the process of PANAMA.

PCAKMX In this approach, we substitute ICA with PCA in estimating a lower dimensional representation. Running PCA will result in n orthogonal components of g dimensions $\mathbf{P}$, and the relative amount of the total variance each component explains. These components can project the measured expression levels $\mathbf{Y}$ onto a lower dimensional space $\mathbf{T}$.

$$\mathbf{YP^T} = \mathbf{T} \tag{4.4}$$

These projections are then tested for associations with genotype values to remove candidate genetic effects, analogous to the process of CONFETI. Then the projections are weighted by the percent variance they explain and a sample covariance matrix is calculated by taking the covariance matrix of the weighted projections.

For PEER we used the `glmApply()` function in the R package `lrgpr` to fit a linear fixed effect model for each phenotype and genotype combination with PEER weights included as covariates. Linear mixed models for CONFETI-I, CONFETI-P, PANAMA, PCAKMX, and ICE were fit using the `lrgpr()` function from the same package.

4.2.3 Genomic Inflation Factor to Assess Model Fit

To assess model fit and to avoid any systematic inflation or deflation of the p-values, we calculated the genomic inflation factor λ statistic for each expression phenotype. The λ statistic was calculated per gene using the median p-value m_p

as

$$\lambda_p = \text{qchisq}(1 - m_p)/\text{qchisq}(0.5) \tag{4.5}$$

where qchisq is a quantile function for the chi-square distribution with 1 degree of freedom. For each method we assessed inflation using λ_p values for every gene to calculate $\lambda_{\text{diff},p} = 1 - \lambda_p$.

4.2.4 Performance Evalulation

To evaluate performance for each method, we ranked the eQTL for each method according to their p-values and then calculated the True Positive Rate (TPR) and False Positive Rate (FPR) and generated Receiver Operating Characteristic (ROC) curves for each method, where we also calculated the area under the curve for each method across the simulation scenarios. True eQTL were further labeled as *cis*, *trans* or broad impact and the recovery rates for each category at different FDR thresholds were calculated by dividing the number of true genotype phenotype pairs that were called significant by the total number of true genotype phenotype pairs in each category. To provide an upper bound metric on how well methods could recover each of these eQTL types, we also simulated the same scenarios without any confounding factors and reported the ROC curves after running LINEAR. We labeled these results 'TMR' for 'Theoretical Maximum Recovery' since these represent the maximum recovery expected in theory if non-genetic factors were perfectly modeled by the confounding factor methods.

4.3 Results

In our analysis of simulated data, we assessed the performance of the eQTL analysis methods CONFETI-I, CONFETI-P, PANAMA, PCAKMX, ICE, PEER, and LINEAR on their ability to identify three types of eQTL, *cis*, *trans* and broad impact, in the presence of confounding factors. We also included the theoretical maximum recovery (TMR) as an upper limit of eQTL detection for each eQTL category, where the phenotype data has only normally distributed random noise added without any confounding factor effects.

4.3.1 Model Fit

First, we evaluated the model fit by calculating genomic inflation factors for each method to detect any significant inflation or deflation of p-values. In the presence of sparse confounding factor effects, CONFETI-I, CONFETI-P, PCAKMX, PEER, and LINEAR showed a moderate level of p-value inflation (Figure 4.1), while PANAMA and ICE showed slightly more conservative model fits compared to other methods. The same trend was observed in the presence of mixed confounding factor effects, with the only difference being that PCAKMX had a slightly less inflated model fit in the latter case. Interestingly, TMR showed a moderate inflation of p-values as well without any confounding factor effects present. This led to the conclusion that the inflation of p-values were likely caused by the large LD blocks observed in the yeast genome, which result in lower p-values in many genotypes due to their correlation with

the causal variant. The more conservative fit observed in PCAKMX, ICE, and PANAMA in comparison to TMR, could be explained by the methods incorrectly explaining away genetic effects as confounding variance.

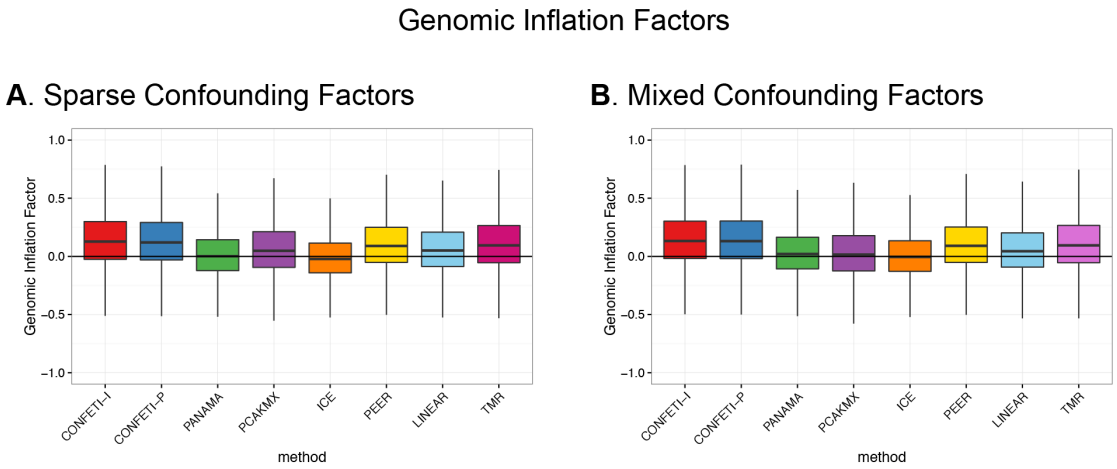


Figure 4.1: Model fit assessment for simulated data

Box-plots of genomic inflation factors calculated for each method across the 50 simulated datasets with **A.** sparse and **B.** mix of sparse and dense confounding factors.

4.3.2 Overall Method Performance Comparison

For both sparse and dense confounding factor effects, all confounding factor correction methods showed a significant increase in the AUC of ROC curves over LINEAR (linear regression without confounding factor correction). Out of the compared confounding factor correction methods, CONFETI-I, closely followed by CONFETI-P, showed the largest improvement over LINEAR in both scenarios of sparse and mixed confounding factors. PANAMA, ICE and PEER showed similar degrees of improvement. Interestingly, PCAKMX showed the lowest degree of improvement in both scenarios. We first hypothesized that

this occurred due to the incorrect removal of components associated with tested genotypes, since PCA was likely to estimated composite effects. However, upon closer inspection we found that both the number of removed principal components and their corresponding contribution to the overall variance were not significant, thus we concluded that constructing the covariance matrix based on PC projections weighted by the percent variance they explain was a sub-optimal strategy.

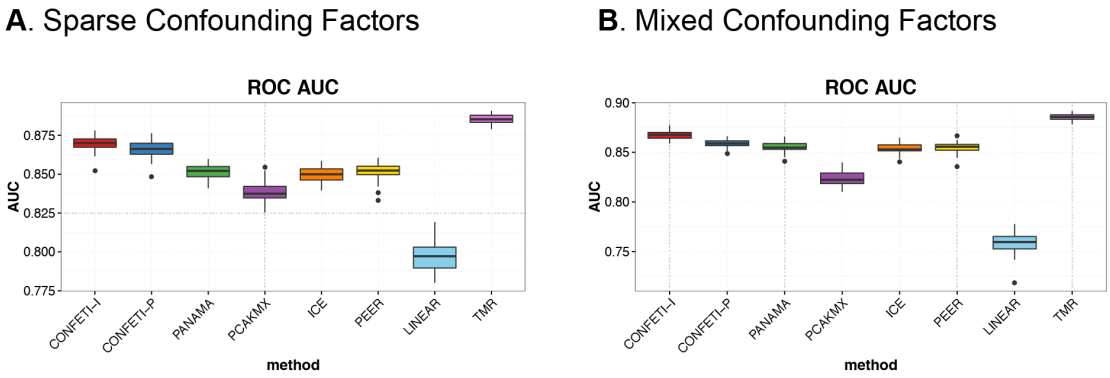


Figure 4.2: Area Under Curve (AUC) calculated for ROC curves

The Area Under the Curve (AUC) for the receiver operator characteristic (ROC) curves for simulated data with **A.** sparse confounding factors and **B.** mix of sparse and dense confounding factors.

4.3.3 True Positive Rate by Simulated eQTL Categories

To further investigate the accuracy in recovering simulated eQTL, we compared the true positive rate (TPR) of each eQTL category (*cis*, *trans*, and broad impact) at varying FDR thresholds. CONFETI-I and CONFETI-P correctly identified the most eQTL in all three categories in the presence of sparse confounding factors (Figure 4.3**A**). For broad impact eQTL in particular, CONFETI-I and CONFETI-

P outperformed all other methods by a large margin. PEER and PCAKMX also performed relatively well in estimating broad impact eQTL over ICE and PANAMA, but fell short in identifying individual *cis*- and *trans*- eQTL. The difference in broad impact eQTL recovery seemed to largely stem from the distinction between genetic and non-genetic effects, since methods that removed genetic effects from the sample covariance matrix (CONFETI-I, CONFETI-P, PCAKMX) seemed to generally perform better than the methods which made no such distinction (PANAMA, ICE). This did not affect the performance of PEER in broad impact eQTL discovery at such extend, since PEER estimates a fixed number of gaussian confounding factors which likely would have not led to an accurate estimate of the broad impact eQTL effects.

The difference between the confounding factor methods decreased with a combination of sparse and dense confounding factors compared to cases with just sparse confounding factors (Figure 4.3**B**), especially in the identification of broad impact eQTL. This is likely due to the difference between the relative amount of total variance explained by each confounding factor and broad impact eQTL. In the dense confounding factor scenario, the confounding factors contribute a significantly higher proportion of the total variance compared to broad impact eQTL. In such a case, distinguishing genetic variance from non-genetic variance has less influence on the covariance matrix estimation, since the majority of the variation in the data is originating from the confounding factors, and the resulting difference between methods in identifying true eQTL is expected to be smaller. Overall, CONFETI-I still identified eQTL in all three categories most accurately, but interestingly CONFETI-P fell below ICE and PANAMA in *cis* and *trans*-eQTL recovery.

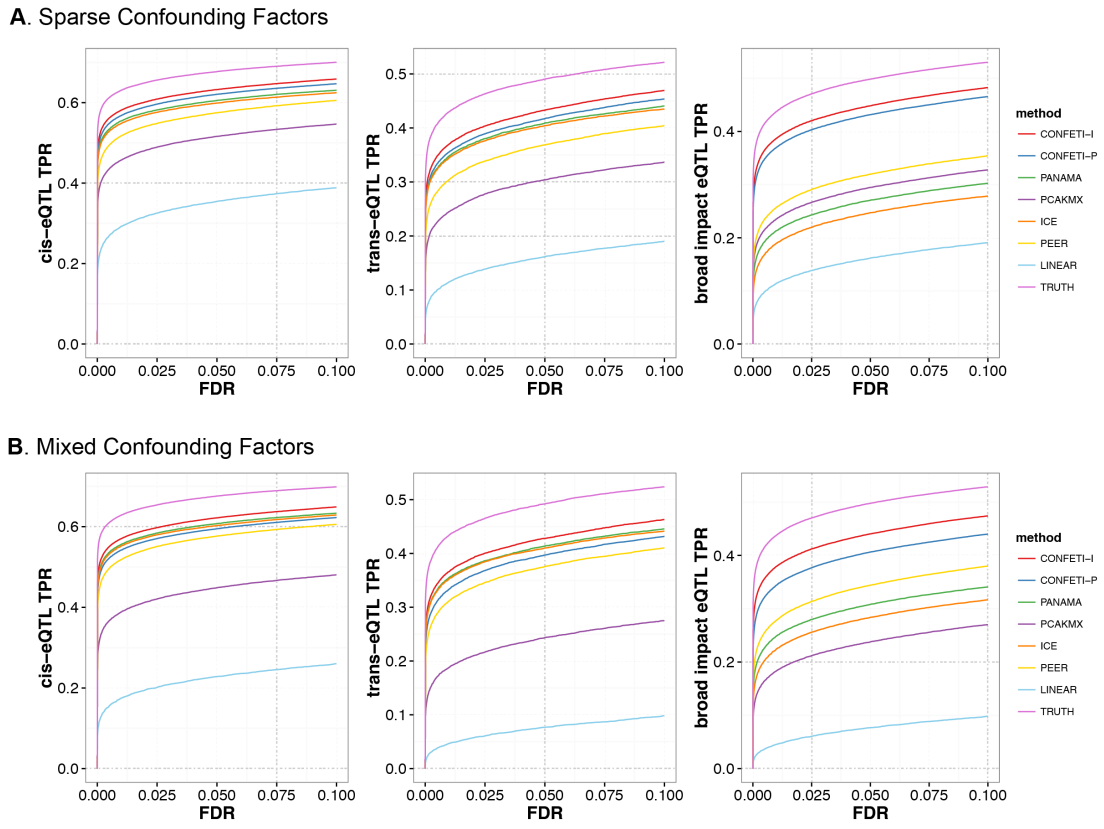


Figure 4.3: True Positive Rate calculated for each simulated eQTL category

The recovery rate of simulated *cis*- (left), *trans*- (middle) and broad impact eQTL (right) for a range of FDR significance thresholds for each method averaging over the 50 simulated datasets with a **A.** sparse and **B.** a mix of dense and sparse confounding factors. The theoretical maximum recovery (TMR) shows the recovery when no confounding factors are included.

4.4 Conclusion

In summary, we demonstrated that linear mixed models that correctly modeled the sample structure with their covariance matrices more accurately detected *cis* and *trans*-eQTL compared to other methods. However, approaches such as ICE and PANAMA, which do not explicitly remove genetic effects from their covariance matrix construction, incorrectly modeled broad impact eQTL as confound-

ing factors. While the extent to which any simulated data will capture the true confounding factor conditions and genetic architectures of real eQTL datasets is unknown, these simulations demonstrate that the CONFETI framework can provide a considerable performance improvement compared to mixed model confounding factor methods in some situations, and performed at least as well as other methods overall.

CHAPTER 5

EVALUATING THE PERFORMANCE OF CONFOUNDING FACTOR CORRECTION METHODS THROUGH REPLICATING EQTL IN HUMAN DATA

While an increasing number of eQTL results are published every year, human eQTL studies lack a standardized methodology and often have different platforms for genotyping and expression measurements. Since the direct comparison of results between studies is difficult, the validity of an identified eQTL is not known for most cases. Therefore, without experimental validation evaluating the quality of eQTL findings is challenging in human datasets. To evaluate the performance of CONFETI compared to other confounding factor correction methods, we re-analyzed multiple human datasets using a central pipeline and inspected the replication of eQTL findings between datasets. We note that this is an imperfect metric that will tend to undercount true positives, however, replication does provide relative control over non-systematic false positives, such that a method that is overly liberal in calling eQTL false positives will be appropriately assessed.

The ideal way to assess the replication of eQTL would be to use exact biological replicates, however, since such datasets are not readily available we first analyzed datasets from the Multiple Tissue Human Expression Resource (MuTHER), which consisted of a cohort of 856 monozygotic and dizygotic twins from the TwinsUK adult registry [18]. We split the dataset by twin pairs to closely mimic biological replicates resulting in subsets with highly similar genetic variation structures. In summary, we analyzed a total of 6 subsets derived

from adipose, lymphoblastoid cell lines (LCL), and skin datasets.

To investigate and compare the analysis results of the MuTHER dataset, we expanded our scope to additional tissue types by analyzing data from the Genotype-Tissue Expression (GTEx) consortium [22]. GTEx datasets did not have the same twin design as MuTHER, but multiple tissue samples were taken from a single individual resulting in similar genetic structures across tissues. We selected 4 pairs of samples that originated from similar tissue types with comparable sample sizes to assess the replication of eQTL.

We first evaluated the model fit by calculating the genomic inflation factor statistic for each phenotype in every dataset. Secondly, we investigated the number of identified *cis* and *trans*-eQTL in each individual dataset in both MuTHER and GTEx. We then compared the eQTL findings between twin pairs in the MuTHER analysis, and between similar tissue types in the GTEx analysis. Finally, we inspected replicating *cis* and *trans*-eQTL across all datasets for both MuTHER and GTEx.

5.1 Methods

5.1.1 Analysis of MuTHER Datasets

We ran eQTL analysis on the adipose, LCL, and skin datasets obtained through the MuTHER project. Based on the matched twins information, there were 161 monozygotic and 220 dizygotic twin pairs in the dataset. We only selected sam-

ples that had both genotype and gene expression measurements for both individuals in each twin pair for all three tissue types. Then we split each tissue specific dataset into two subsets separating each twin pair into different subsets. This created two subsets for each tissue type resulting with 327 samples for adipose, 329 for LCL, and 253 samples for skin. From the downloaded genotypes, we used only non-imputed genotypes with minor allele frequencies higher than 0.05.

Table 5.1: **Sample size, number of genes and genotypes for each MuTHER dataset analyzed.**

Tissue	Sample Size	Gene Expression	Genotypes
Adipose	327	28964	246298
LCL	329	28894	246298
Skin	253	28893	246298

5.1.2 Analysis of GTEx Datasets

We selected 4 pairs of tissues (Adipose, Artery, Heart, Skin) from GTEx release v6 (dbGaP Accession phs000424.v6.p1) with over 150 samples that have both RNA-seq gene expression and SNP array genotypes (Table 5.2). For gene expression measurements, we used the pre-processed expression values directly obtained from the GTEx portal. For genotypes, we excluded SNPs with missing genotypes and those with minor allele frequency < 0.05. We also pruned SNPs within 10kb with pairwise $r^2 > 0.99$ and removed SNPs which were deprecated in dbSNP (1,270,565 SNPs remaining).

Table 5.2: **Sample size, number of genes and genotypes for each GTEx dataset analyzed.**

Tissue	Subtype	Sample Size	Gene Expression	Genotypes
Adipose	Subcutaneous	298	27182	1270565
Adipose	Visceral	185	26261	1270565
Artery	Aorta	197	25292	1270565
Artery	Tibial	285	25311	1270565
Heart	Atrial Appendage	159	24541	1270565
Heart	Left Ventricle	190	23710	1270565
Skin	Leg	302	27815	1270565
Skin	Suprapubic	196	26913	1270565

5.1.3 eQTL Analysis

We fit CONFETI-I, CONFETI-P, PANAMA, PCAKMX, ICE, PEER, and LINEAR for every phenotype and genotype pair using the method settings and parameters as described in the previous chapter. We used genotype principal components included as covariates (2 for each MuTHER and 3 for GTEx datasets) to account for population structure. Additionally, we used age and experimental batch as covariates for the MuTHER analysis, and used gender and genotyping platform as covariates for the GTEx analysis. After calculating p-values for all phenotype and genotype pairs, we adjusted the p-values using Benjamini-Hochberg multiple hypothesis correction. The corrected p-values represent upper bounds on False Discovery Rate (FDR) [127]. We used a threshold of 0.01 on the adjusted p-values to mark significant eQTL. An eQTL (significant SNP gene pair) was labeled as *cis* if the SNP and gene were located on the same chromosome within 1 Mb, and *trans* otherwise. We screened all *trans*-eQTL for cases where the SNP was coincident either with an annotated gene copy (such as a pseudogene or functional gene 'parent' of a pseudogene), or a region of

the genome with high sequence similarity (covering at least 30% of the gene transcript) to the eQTL gene. These unannotated regions were identified by aligning all gene transcripts to the entire genome using the BLAT tool [128]. This "pseudo-*trans*" screening revealed that a number of the replicating *trans*-eQTL were artifacts arising due to incorrect/ambiguous mapping of RNA-seq reads that are in fact caused by *cis*-regulation of a gene which shares sequence similarity with the eQTL gene. In order to avoid double-counting eQTL associated with multiple linked SNPs, we selected at most one significant *cis*- and *trans*- SNP per cytoband per gene. Using these criteria, we measured the replication of eQTL between and across different tissues counting the overlapping cytoband and gene pairs that were called significant in each dataset.

5.2 Results

5.2.1 Model fit

We ran each of the eQTL analysis methods on six MuTHER [18] datasets consisting of three twin pairs for adipose, LCL and skin samples and the eight GTEx [22] datasets made up of four tissue pairs (Adipose, Visceral vs. Subcutaneous; Artery, Aorta vs Tibial Artery; Heart, Atrial Appendage vs. Left Ventricle; Skin, Leg vs. Suprapubic). For each method applied to each dataset, we inspected the median λ genomic inflation factor [79] as a measure of appropriate model fit and control of false positives and false negatives rates.

MuTHER All methods were within acceptable fit levels with no significant inflation or deflation with genotype PCs included as covariates (Figure 5.1**A**). One interesting observation was that ICE consistently showed a slightly higher measure of inflation factors in all datasets compared to other methods. This was not observed in the analysis of simulated data, where CONFETI-I and CONFETI-P showed a slight inflation. This could be caused by an increased number of false positives, or it could illustrate that ICE corrects the confounding factor effects best and is able to recover the highest number of eQTL results. However, without the validation of results returned by all methods, the exact cause remains unknown. Additionally, an overall trend of linear fixed effect models showing a slightly more conservative fit than linear mixed model based approaches was observed in all datasets (Figure 5.1**A** and Figure B.1**A**).

GTEx Similar to the model fit evaluation of MuTHER datasets, we observed a slight inflation of p-values in ICE results (Figure 5.1**B**, Figure B.1**B**). Linear fixed effect models also tended to be more conservative than linear mixed models in the GTEx data, and the largest difference between methods was observed in the GTEx heart atrial appendage dataset (Figure 5.1**B**). The heart atrial appendage dataset showed also the highest degree of inflation in all of the linear mixed model approaches. Interestingly, this dataset had the smallest sample size, which could have resulted in a less accurate estimation of the sample structure by confounding factor methods (Table 5.2). This observation is also repeated in the skin subsets of MuTHER (Figure 5.1**A**, Figure B.1**A**) with the smallest sample size out of the three tissues (Table 5.1).

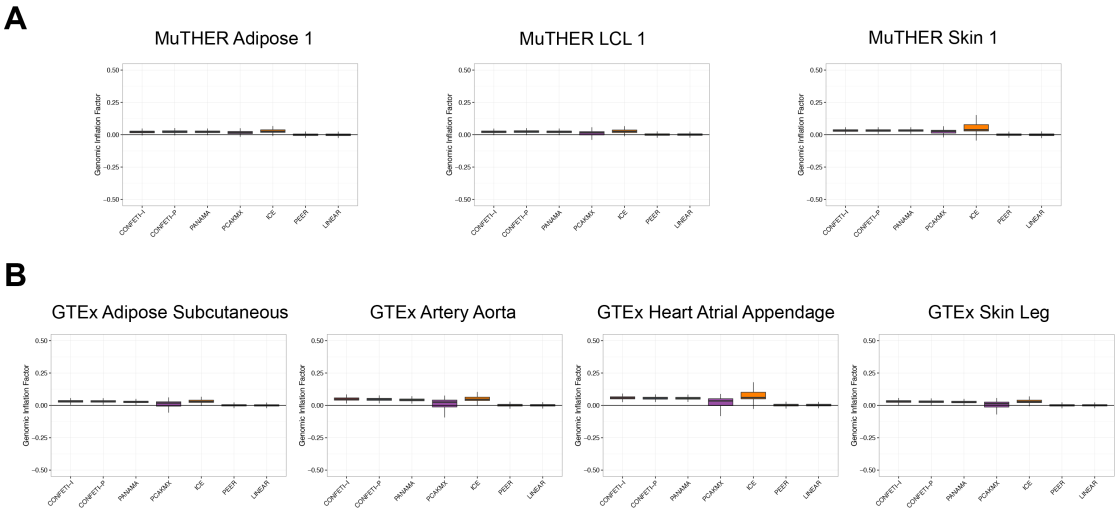


Figure 5.1: MuTHER and GTEx model fit evaluation

Genomic inflation factors calculated for each analysis method in each sub-dataset of **A.** MuTHER and **B.** GTEx.

5.2.2 eQTL Discovery in Individual Datasets

MuTHER The number of both *cis* and *trans*-eQTL discoveries increased for less stringent FDR thresholds in all of the datasets (Figure B.3). A significant difference between *cis* and *trans*-eQTL was that while the number of *trans*-eQTL increased steadily with a linear relationship to the significant threshold, the count of *cis*-eQTL seemed to increase at a much lower rate and most *cis*-eQTL were identified even with the most stringent threshold. This could be explained by the generally lower effect sizes of *trans*-eQTL, but could also be caused by an increase of false positive discoveries as the threshold is lowered.

Confounding factor methods greatly increased the number of identified *cis*-eQTL in all datasets by approximately 2-3 fold compared to LINEAR, demonstrating the increase of power by accounting for systematic variation. Little dif-

ference was found in the number of identified *cis*-eQTL between linear mixed model confounding factor methods, except for PCAKMX which found notably fewer *cis*-eQTL in all datasets compared to CONFETI-I, CONFETI-P, PANAMA and ICE. In agreement with the simulation results (Figure 4.2), PEER identified more *cis*-eQTL in both the adipose and LCL dataset compared to PCAKMX and a comparable number of *cis*-eQTL in the skin dataset, illustrating that confounding factor estimation through PEER is a better strategy for *cis*-eQTL discovery in comparison to PCAKMX.

Considering an FDR threshold of 0.01, fewer *trans*-eQTL were identified compared to the number of cis-eQTL in all datasets and all methods. ICE found the most *trans*-eQTL in all cases except the second subset of the skin dataset, and similar to *cis*-eQTL the observed difference between linear mixed model based approaches was minimal. The difference in *trans*-eQTL discovery in the second subset of the skin dataset seems to be driven from the removal of components associated with genotypes, since only the methods which removed candidate genetic effects (CONFETI-I, CONFETI-P, and PCAKMX) show an increase in *trans*-eQTL identification in comparison to the first subset. Another interesting observation was that out of the confounding factor correction methods, only PEER recovered fewer *trans*-eQTL in comparison to LINEAR, in one subset of adipose and in both skin subsets compared to LINEAR. This could be potentially explained by the lower statistical power of PEER in detecting weaker effects due to the increased number of parameters, but could also depict the potential over correction of real effects by PEER.

GTEx Results obtained by analyzing tissue pairs from the GTEx dataset presented similar findings to the MuTHER dataset. *cis*-eQTL discovery started to asymptote while the number of identified *trans*-eQTL steadily increased in all datasets and all analysis methods (Figure B.4). The absolute number of identified eQTL differed between datasets showing a positive correlation with larger sample size (Figure B.2).

Consistent with the findings in the MuTHER dataset, confounding factor methods greatly increased the number of identified *cis*-eQTL compared to LINEAR with the largest increases observed with linear mixed model based methods ICE, CONFETI-I, CONFETI-P, and PANAMA. Out of the linear mixed model based methods, ICE constantly identified the most *cis*-eQTL across all datasets. PEER also found more *cis*-eQTL in comparison to PCAKMX in all datasets except the heart artial appendage dataset, reinforcing the finding in MuTHER that given a sufficient sample size PEER outperforms PCAKMX in *cis*-eQTL discovery.

ICE identified the most *trans*-eQTL in all analyzed datasets, followed by CONFETI-I, CONFETI-P and PANAMA, which showed minimal difference in the number of identified *trans*-eQTL. Unlike *cis*-eQTL PEER constantly found fewer *trans*-eQTL compared to PCAKMX, but identified more hits compared to LINEAR.

In summary, correction for confounding factor effects increased the number of identified *cis*-eQTL in all datasets when compared to the linear model with only known covariates included. The greatest increase in identified hits

was observed when a linear mixed model which accounted for most of the observed variance was used (CONFETI-I, CONFETI-P, ICE, and PANAMA). Similar trends were observed in *trans*-eQTL discovery as well, with a few exceptions of PEER finding fewer *trans*-eQTL in comparison to LINEAR in the analysis of MuTHER datasets.

5.2.3 Replicating eQTL Compared Across Methods

To further investigate the results generated by all methods in both MuTHER and GTEx datasets, we first assessed the replication of each identified eQTL between each twin pair in MuTHER and in each tissue pair dataset in GTEx.

MuTHER First, in Figure 5.2**A** we show the replication of *cis*-eQTL in the adipose twin-pair datasets. The identification of replicating *cis*-eQTL at different FDR thresholds resembled the discovery of *cis*-eQTL in each individual dataset. With most of the replicating *cis*-eQTL being identified at very stringent significance thresholds, the number of replicating *cis*-eQTL increased at a low rate for less stringent thresholds. For an FDR threshold of < 0.01, the replicating *cis*-eQTL were largely overlapping between CONFETI-I, CONFETI-P, PANAMA and ICE, while only a few unique results were found by each method. Additionally, replicating *cis*-eQTL identified by PCAKMX, PEER and LINEAR were subsets of hits identified by other methods, showing that CONFETI-I, CONFETI-P, PANAMA, and ICE were identifying eQTL additional to the results of LINEAR, PEER, and PCAKMX. The same trend was observed in the analysis of LCL (Fig-

ure B.5**A**) and in skin (Figure B.6**A**) datasets. For each analysis method in each MuTHER subset pair, the percentage of replicating *cis*-eQTL were between 60% to 80% of *cis*-eQTL (Figure 5.3**A**).

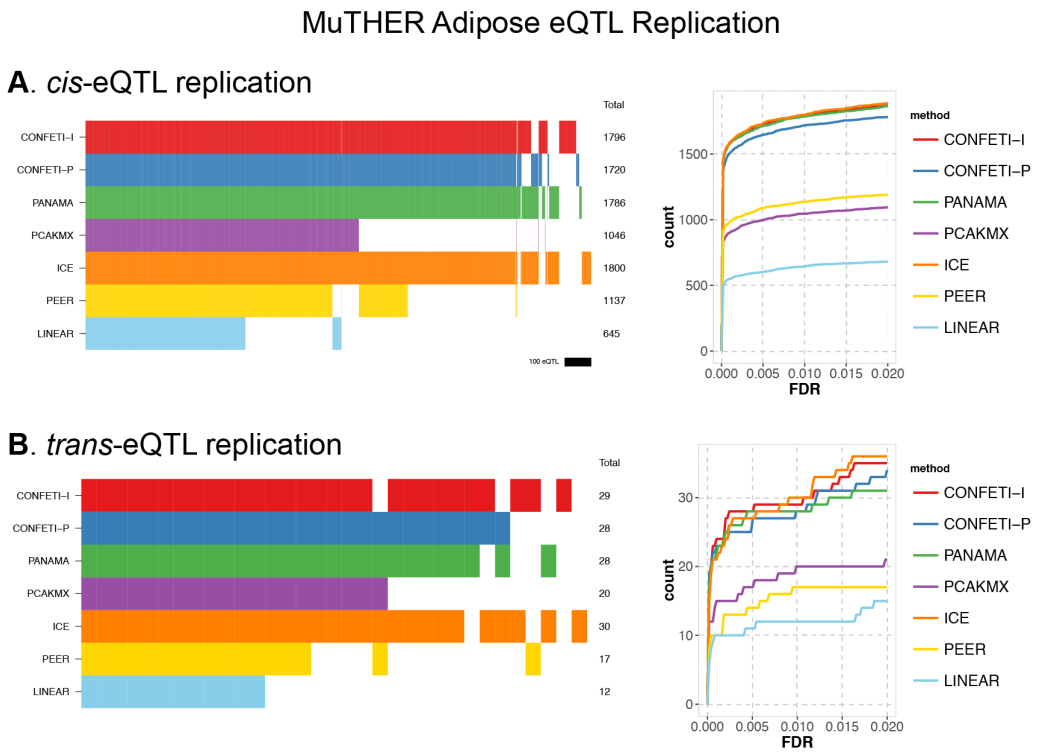


Figure 5.2: Replicating eQTL in the MuTHER Adipose Subsets

Plots showing replicating eQTL for **A.***cis* and **B.***trans*-eQTL at an FDR threshold of 0.01 across all methods (left) and the count of replicating eQTL versus various FDR thresholds for each of the analysis methods (right). In the plots on left, replicating eQTL are ordered on the x-axis by the amount of overlap between methods. Colored bars corresponding to their method indicate that the particular eQTL replicated, and the total number of replicating eQTL for each method is shown at the end of each bar.

In contrary to *cis*-eQTL, we found that a only a very small number of *trans*-eQTL replicated within the adipose twin pair dataset (Figure 5.2**B**). The number of replicating *trans*-eQTL only accounted for less than 10% of the union of unique *trans*-eQTL identified in each dataset with the exception of PEER and

LINEAR in the LCL analysis (Figure 5.3**B**). This trend could be explained by the comparably lower number of *trans*-eQTL that LINEAR and PEER identified, which led to a higher replicating fraction driven by a few replicating eQTL. Moreover, if a few *cis*-eQTL were mis-classified as *trans*-eQTL due to having a distance slightly larger than the threshold of 1Mb or due to mis-mapping of the gene, LINEAR and PEER would likely only identify these as *trans*-eQTL thus having a higher replication rate. Another observed trend was that unlike the sharply increasing *trans*-eQTL findings in each dataset, the number of replicating *trans*-eQTL only showed a slight increase as the significance threshold was lowered. The analysis results of LCL (Figure B.5**B**) and skin (Figure B.6**B**) datasets showed similar findings, with LCL having the most replicating *trans*-eQTL. Based on these results, we concluded that the low replication rate of *trans*-eQTL cannot be solely attributed to lower effect size, which raised concerns about their credibility as biological findings.

GTEx To present a direct comparison with the MuTHER analysis, in Figure 5.4**A** we first present the results for replicating *cis*-eQTL identified in the GTEx adipose tissue pair. We found a large number of replicating *cis*-eQTL between the GTEx adipose subcutaneous and visceral datasets, and the majority of replicating *cis*-eQTL were identified at stringent thresholds. CONFETI-I, CONFETI-P, PANAMA, and ICE found similar numbers of replicating *cis*-eQTL with ICE having the most unique replicating eQTL. Interestingly, the number of replicating *cis*-eQTL found by PEER and PCAKMX were comparable, despite the fact that PEER identified more *cis*-eQTL hits in each dataset separately. Ranging between 35% to 45% the percentage of replicating *cis*-eQTL was lower

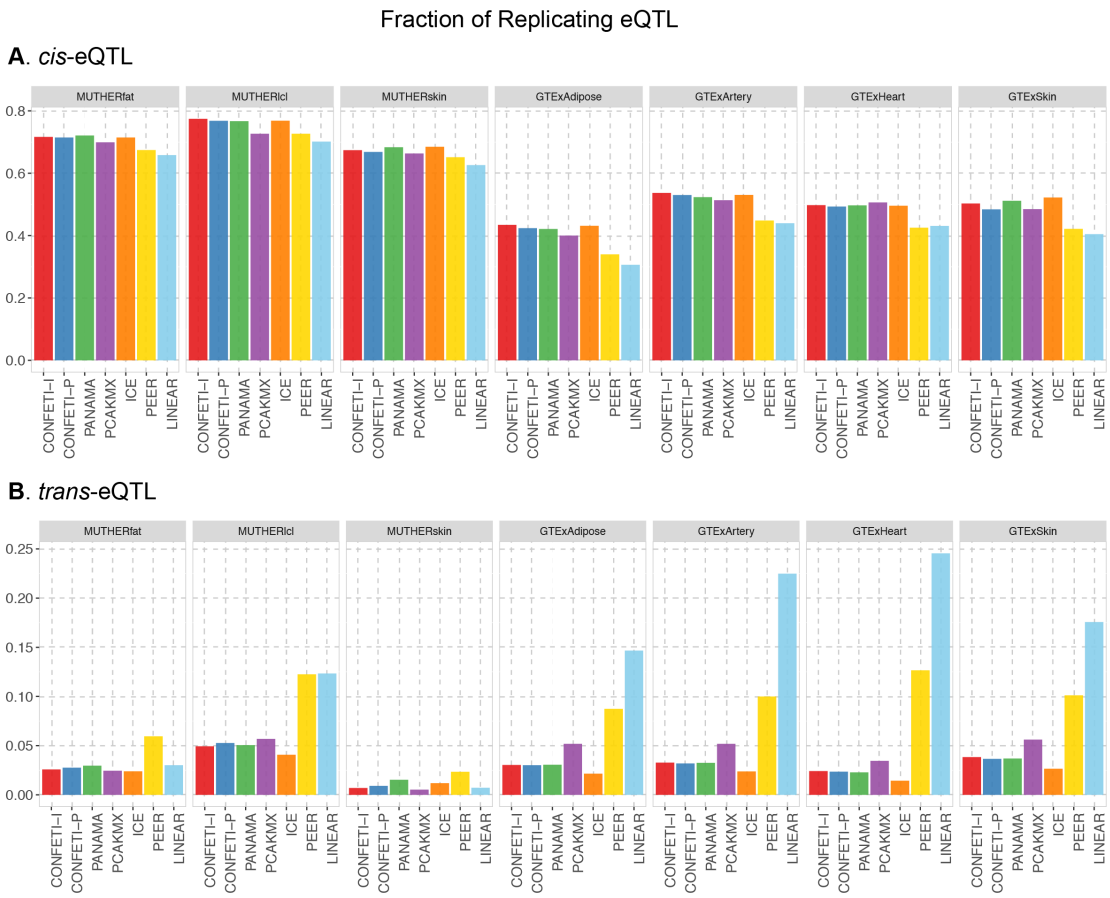


Figure 5.3: Fraction of replicating eQTL in the MuTHER and GTEx datasets

A. Fraction of replicating *cis*-eQTL and **B.** *trans*-eQTL are calculated by dividing the number of replicating eQTL by the union of unique eQTL found in each dataset pair.

compared to the MuTHER analysis (Figure 5.3**A**). This could be due to the difference between the tissue subtypes which was a heterogeneity not present in the MuTHER dataset. Another explanation could be the difference in sample size between subcutaneous (298) and visceral (185) subsets, leading to more *cis*-eQTL overall which are primarily identified in the subcutaneous dataset. However, given the higher replication ratio in the artery dataset, which has a similar sample size difference to the adipose subsets (Aorta 197 vs Tibial 285), the dif-

ference between the tissue types would be a more likely explanation.

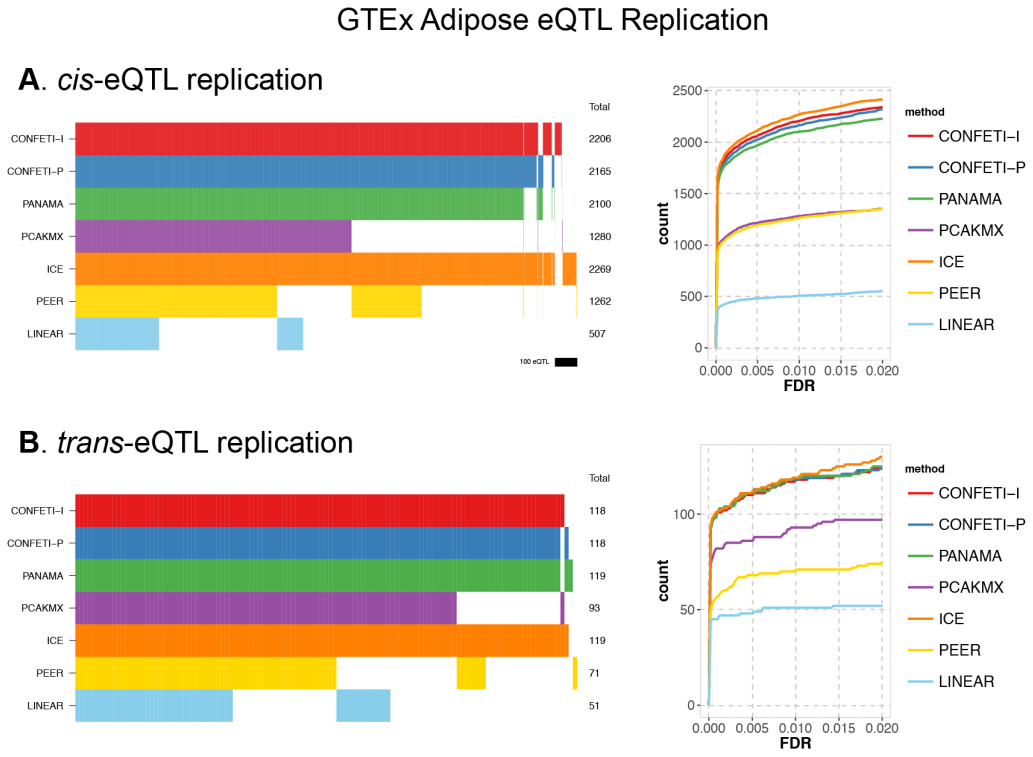


Figure 5.4: Replicating eQTL between GTEx Adipose Subcutaneous and Visceral

Plots showing replicating eQTL for **A.***cis* and **B.***trans*-eQTL at an FDR threshold of 0.01 across all methods (left) and the count of replicating eQTL versus various FDR thresholds for each of the analysis methods (right). In the plots on left, replicating eQTL are ordered on the x-axis by the amount of overlap between methods. Colored bars corresponding to their method indicate that the particular eQTL replicated, and the total number of replicating eQTL for each method is shown at the end of each bar.

A higher number of replicating *trans*-eQTL were identified in the GTEx adipose analysis compared to the MuTHER adipose results (Figure 5.4**B**), however, *trans*-eQTL still showed poor replication rates when considering the overall number of identified *trans*-eQTL in each dataset (Figure 5.3**B**). *trans*-eQTL identified by LINEAR and PEER showed a much higher replication ratio compared to other methods, which could be explained by similar reasons mentioned

in the MuTHER replication analysis. The replicating *cis* and *trans*-eQTL results of artery (Figure B.7), heart (Figure B.8), and skin (Figure B.9) datasets showed consistent trends in comparison to the adipose dataset results.

In summary, we demonstrated that in both matched twin-pair and tissue-pair analysis of multiple tissues that *cis*-eQTL findings were highly replicable, whereas *trans*-eQTL showed a much lower replication rate. Results between linear mixed model methods CONFETI-I, CONFETI-P, PANAMA, and ICE showed only small differences identifying largely the same replicating eQTL for both *cis* and *trans*. These methods also found almost all eQTL identified by PCAKMX, PEER, and LINEAR indicating a higher statistical power for methods with estimated sample covariance matrices that account for confounding variance.

5.2.4 Replication of *cis* and *trans* eQTL across datasets

MuTHER Comparing *cis*-eQTL and *trans*-eQTL that were identified across datasets revealed another striking difference between the two categories. *cis*-eQTL showed a high rate of replication between all datasets, with more than 70% of the total *cis*-eQTL being replicated in more than one dataset for all analysis methods (Figure 5.5). Interestingly, the fraction of replicating *cis*-eQTL were broadly similar across all analysis methods. A high fraction of *cis*-eQTL replicated in 2 subsets, which demonstrated the difference in gene regulation between the adipose, LCL, skin tissues. These could be seen as replicating *cis*-eQTL blocks specific to each tissue type in Figure 5.6**A**. Out of the three tissue

types, LCL subsets showed the largest amount of replicating tissue specific *cis*-eQTL (Figure 5.6**A**).

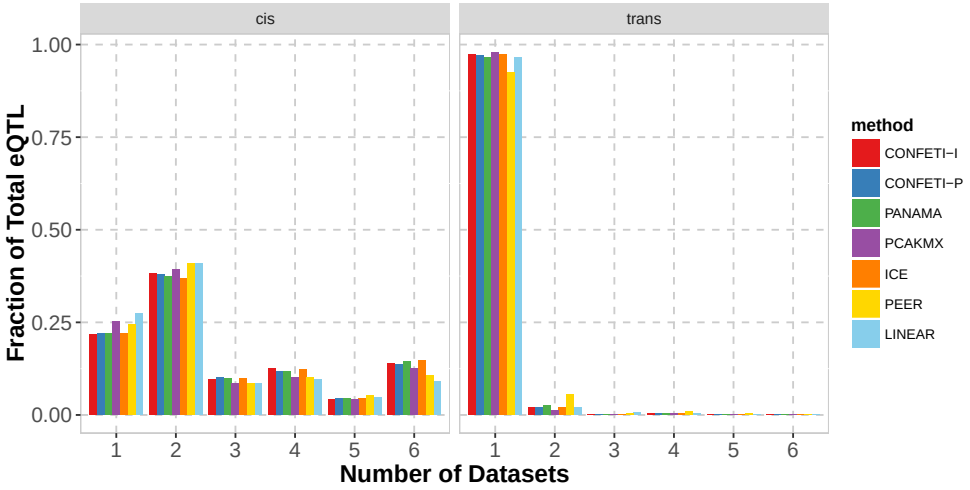


Figure 5.5: Fraction of replicating eQTL by number of MuTHER datasets identified in

Bar plots showing the fraction of total *cis* and *trans*-eQTL by the number of datasets they were found in for the MuTHER analysis.

The replicating fraction of eQTL were also consistent between analysis methods for *trans*-eQTL. However, in contrast to *cis*-eQTL, more than 90% of the *trans*-eQTL were only identified in a single subset showing minimal overlap even between the same tissue types (Figure 5.5). Given the twin pair design of the analysis, we expected to see an enrichment of replicating *trans*-eQTL found in 2 datasets similar to *cis*-eQTL, however the evidence for such a trend was minimal. We observed replicating *trans*-eQTL that were specific to the LCL subsets, however, compared to tissue specific *cis*-eQTL the fraction of such replicating *trans*-eQTL fell short by a large margin (Figure 5.6**B**).

The comparison of replicating eQTL identified by each method across all datasets yielded similar results to the replication results in twin pairs for each

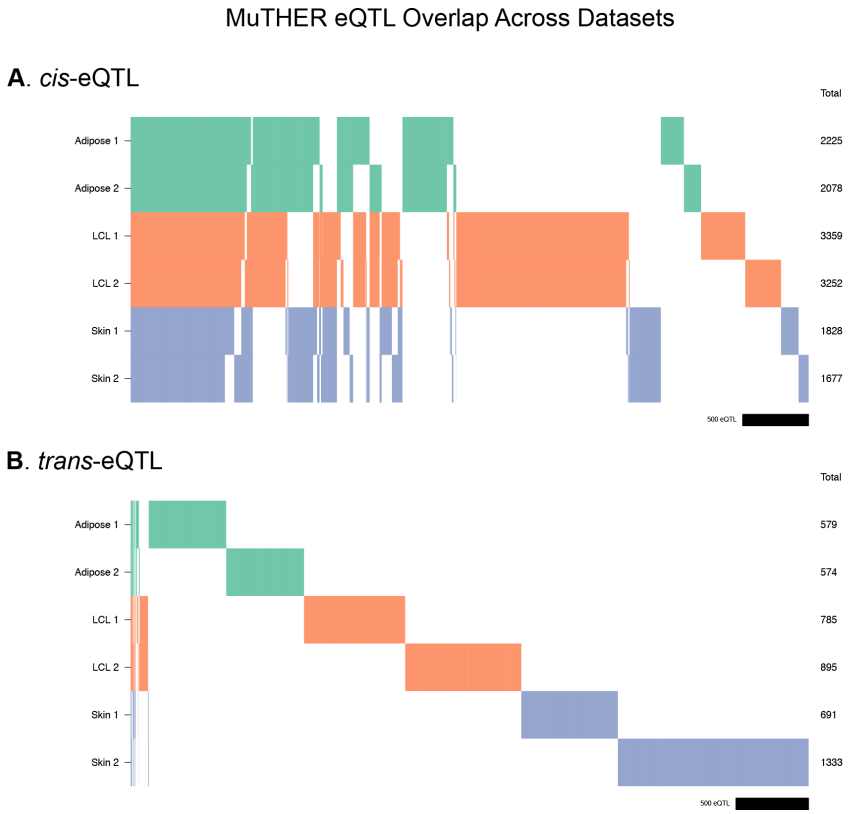


Figure 5.6: *cis*-eQTL and *trans*-eQTL overlap across datasets in the MuTHER analysis

Unique **A.** *cis*-eQTL and **B.** *trans*-eQTL found in each dataset by CONFETI-I are ordered on the x-axis by the amount of overlap between different datasets. Colored bars corresponding to their tissue types (Adipose, LCL, Skin) indicate that the particular eQTL was identified in the dataset, and white space indicates that the eQTL was not found. The total number of replicating eQTL are shown on the right. The black scale on the bottom right of each plot shows the length covering 500 eQTL.

tissue type. ICE found the highest number of *cis*-eQTL replicating across all datasets (756), closely followed by PANAMA (737), CONFETI-I (714), and CONFETI-P (689) at an FDR threshold of 0.01 (Figure 5.7). For *trans*-eQTL, ICE and PANAMA found the same 5 replicating hits across datasets, while others found fewer. Upon closer inspection only 1 *trans*-eQTL out of the 5 had a genotype and gene that were on different chromosomes, while the other 4 were likely

driven by *cis*-eQTL that were slightly outside of the 1Mb window (Figure B.10).

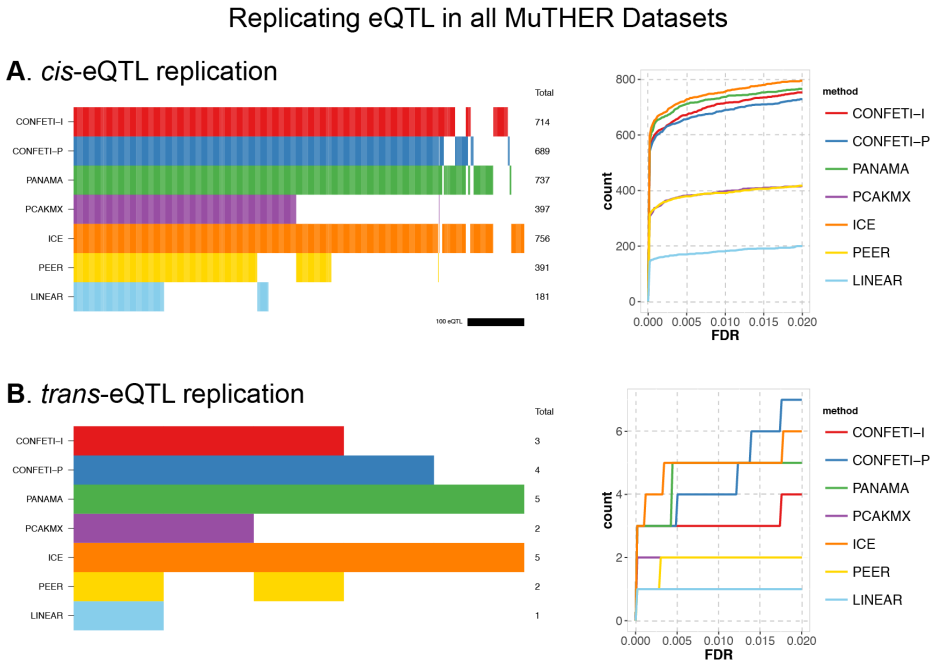


Figure 5.7: Replicating eQTL across all MuTHER datasets

Plots showing replicating eQTL for **A.***cis* and **B.***trans*-eQTL at an FDR threshold of 0.01 across all methods (left) and the count of replicating eQTL versus various FDR thresholds for each of the analysis methods (right). In the plots on left, replicating eQTL are ordered on the x-axis by the amount of overlap between methods. Colored bars corresponding to their method indicate that the particular eQTL replicated, and the total number of replicating eQTL for each method is shown at the end of each bar.

GTEx When we investigated replicating eQTL across all GTEx datasets we found that while more than 60% of the *cis*-eQTL replicated in more than one dataset, the fraction of *cis*-eQTL that replicated between 2 datasets significantly reduced compared to the MuTHER analysis (Figure 5.8). This stems likely from the difference between tissue subsets, since tissue samples obtained from different locations of the body are expected to be less similar than subsets of samples from the same tissue type. Moreover, the cell type composition of the sam-

pled region might also differ, which can be a source of additional heterogeneity. However, we were still able to find tissue type specific *cis*-eQTL in the GTEx analysis, where the overlap between the skin and artery subsets produced the most tissue specific *cis*-eQTL (Figure 5.9).

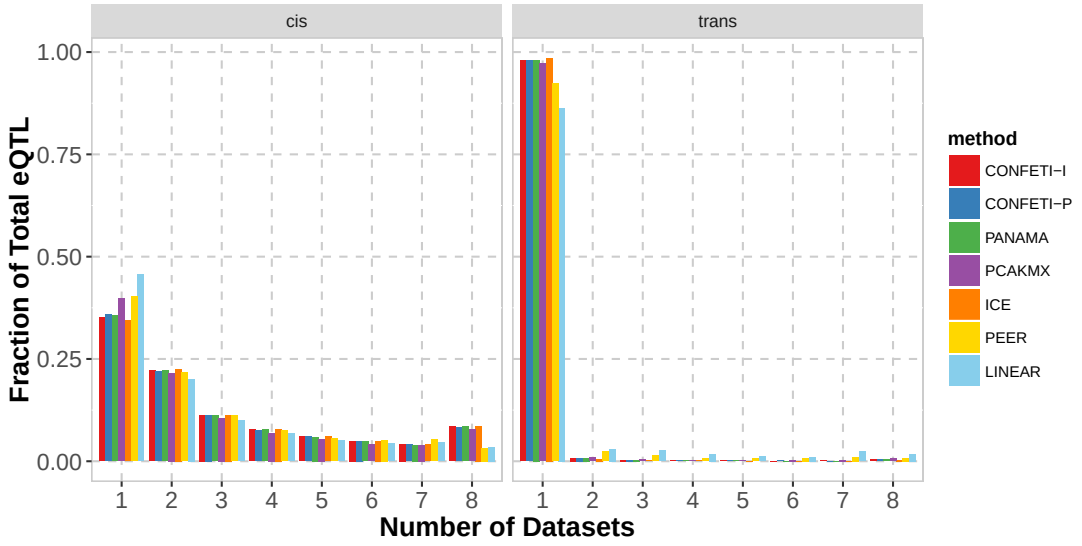


Figure 5.8: Fraction of replicating eQTL by number of GTEx datasets identified in

Bar plots showing the fraction of total *cis* and *trans*-eQTL by the number of datasets they were found in for the GTEx analysis.

The findings in *trans*-eQTL replication across all datasets were consistent with those in the MuTHER analysis, where more than 90% of the identified *trans*-eQTL by all methods except LINEAR were specific to the dataset and showed poor replication overall. LINEAR showed a slightly higher fraction of *trans*-eQTL that replicated in more than one dataset, however this is likely due to the significantly smaller number of *trans*-eQTL identified by LINEAR in comparison to other methods.

At an FDR threshold of 0.01, ICE found the most *cis*-eQTL replicating across

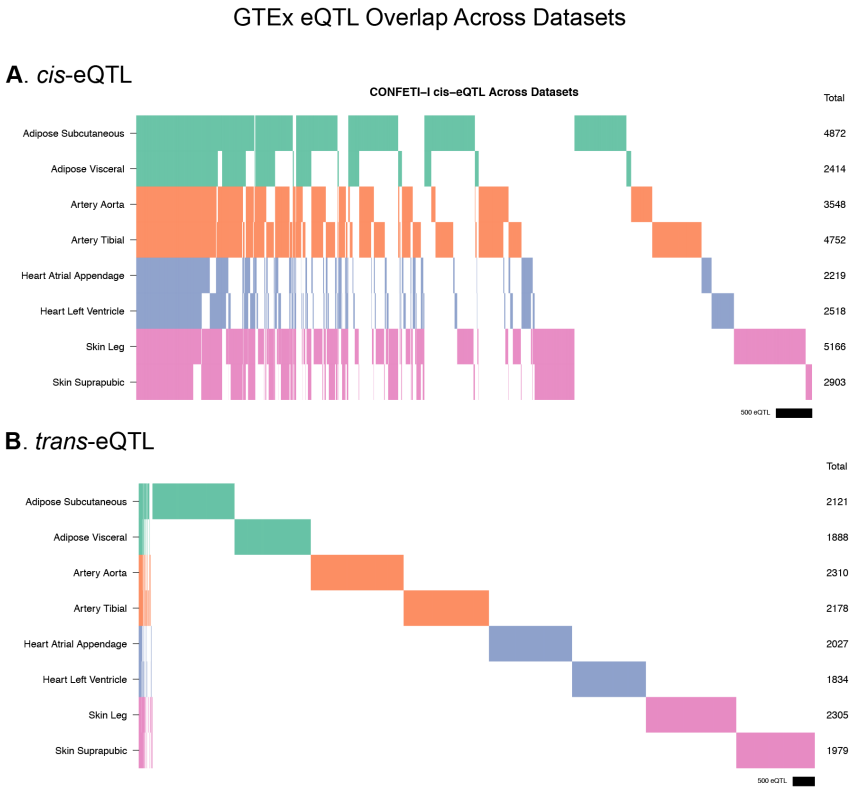


Figure 5.9: *cis*-eQTL and *trans*-eQTL overlap across datasets in the GTEx analysis

Unique **A.** *cis*-eQTL and **B.** *trans*-eQTL found in each dataset by CONFETI-I are ordered on the x-axis by the amount of overlap between different datasets. Colored bars corresponding to their tissue types (Adipose, Artery, Heart, and Skin) indicate that the particular eQTL was identified in the dataset, and white space indicates that the eQTL was not found. The total number of replicating eQTL are shown on the right. The black scale on the bottom right of each plot shows the length covering 500 eQTL.

all datasets (839), and CONFETI-I (796), PANAMA (790), and CONFETI-P (780) produced comparable results. While the number of replicating *trans*-eQTL only showed minimal difference between methods, we found significantly more *trans*-eQTL that replicated in all GTEx datasets. Additionally, unlike the results in MuTHER dataset we found that many of the replicating *trans*-eQTL across datasets consisted of genotype and gene pairs which were positioned on dif-

ferent chromosomes (Figure B.11). However, more than 75% of the replicating *trans*-eQTL were pseudogenes (Figure B.12). Therefore, the increase in *trans*-eQTL replication in GTEx is likely driven by the additional genes measured by RNA-seq and incomplete filtering of pseudogene relationships.

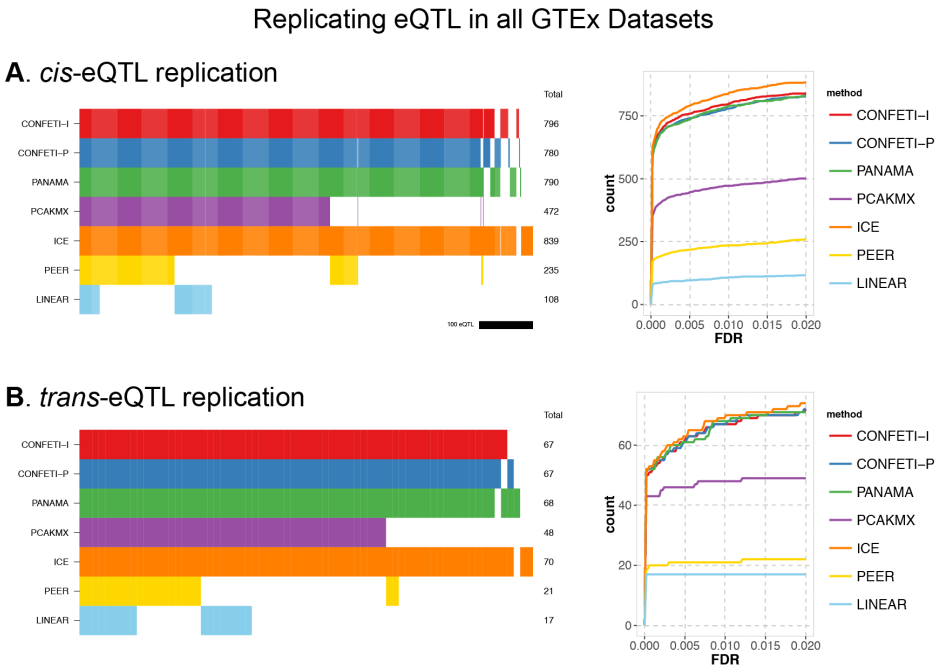


Figure 5.10: Replicating eQTL across all MuTHER datasets

Plots showing replicating eQTL for **A.***cis* and **B.***trans*-eQTL at an FDR threshold of 0.01 across all methods (left) and the count of replicating eQTL versus various FDR thresholds for each of the analysis methods (right). In the plots on left, replicating eQTL are ordered on the x-axis by the amount of overlap between methods. Colored bars corresponding to their method indicate that the particular eQTL replicated, and the total number of replicating eQTL for each method is shown at the end of each bar.

5.3 Conclusion

In this chapter we presented the results of eQTL analysis conducted by applying CONFETI and other confounding factor correction methods to datasets obtained from the MuTHER and GTEx consortium. We found that the application of confounding factor correction methods largely increased the number of identified *cis*-eQTL in all analyzed datasets. Linear mixed model based methods CONFETI-I, CONFETI-P, PANAMA, ICE, and PCAKMX also significantly increased the number of identified *trans*-eQTL. To evaluate the quality of these findings, we investigated replicating eQTL in both *cis* and *trans* categories as performance measures. We found that linear mixed model based confounding factor methods identified *cis*-eQTL that replicated well, with more than 60% replicating in matched twin pairs in the MuTHER analysis, and that more than 40% replicating between similar tissue types in the GTEx analysis. However, despite finding more *trans*-eQTL overall, the replication rate remained below 10% for eQTL identified with linear mixed model confounding factor correction methods. This trend could not be explained solely by smaller effect sizes of *trans*-eQTL, since we would still expect part of the findings to replicate in twin pairs. Thus, we concluded that the *trans*-eQTL findings are likely to be false positives even at a relatively stringent FDR threshold of < 0.01.

Interestingly, we found little difference in both *cis* and *trans*-eQTL discovery and replication between CONFETI-I, CONFETI-P, ICE and PANAMA. To our surprise, the most conservative method ICE identified the largest number of eQTL in most of the cases. This could illustrate that the contribution of genetic effects to the total variance is not significant in human scale data, and account-

ing for the majority of total variance is a good approximation of confounding sample structures.

APPENDIX A

APPENDIX OF CHAPTER 2

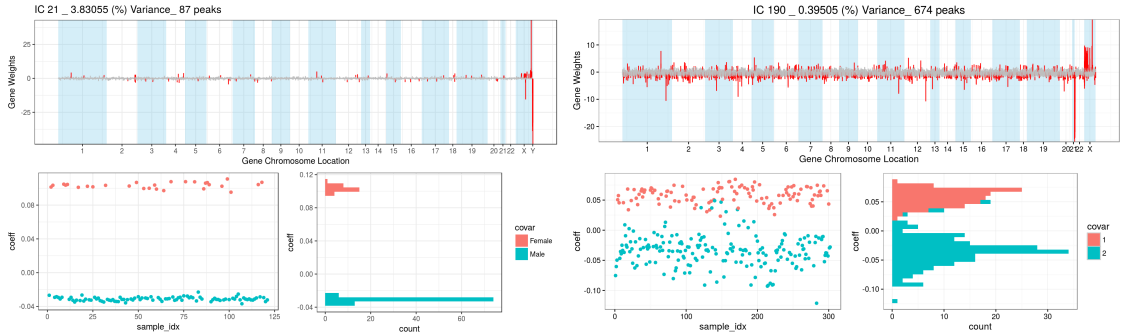


Figure A.1: Additional examples of gender specific expression patterns estimated by ICA

Gender specific expression patterns estimated by ICA in a microarray dataset of smokers and non-smokers (left) and the RNA-seq dataset from GTEx-Skin-Sun-Exposed (right) showing the components with the most significant association with the known covariate gender.

APPENDIX B

APPENDIX OF CHAPTER 5

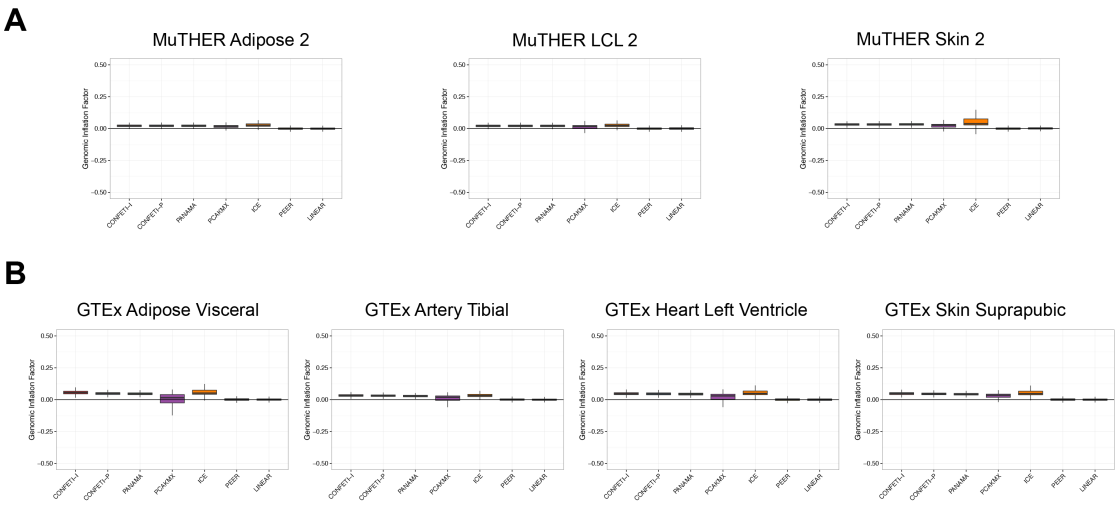


Figure B.1: MuTHER and GTEx model fit 2

Additional genomic inflation factor plots.

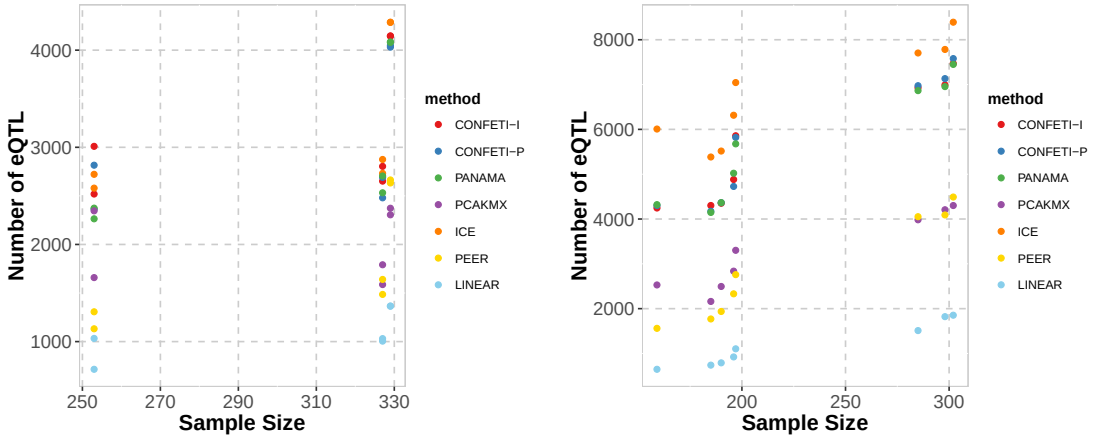


Figure B.2: Total Unique eQTL by Sample Size

The total number of eQTL identified by each method shown by sample size for MuTHER and GTEx.

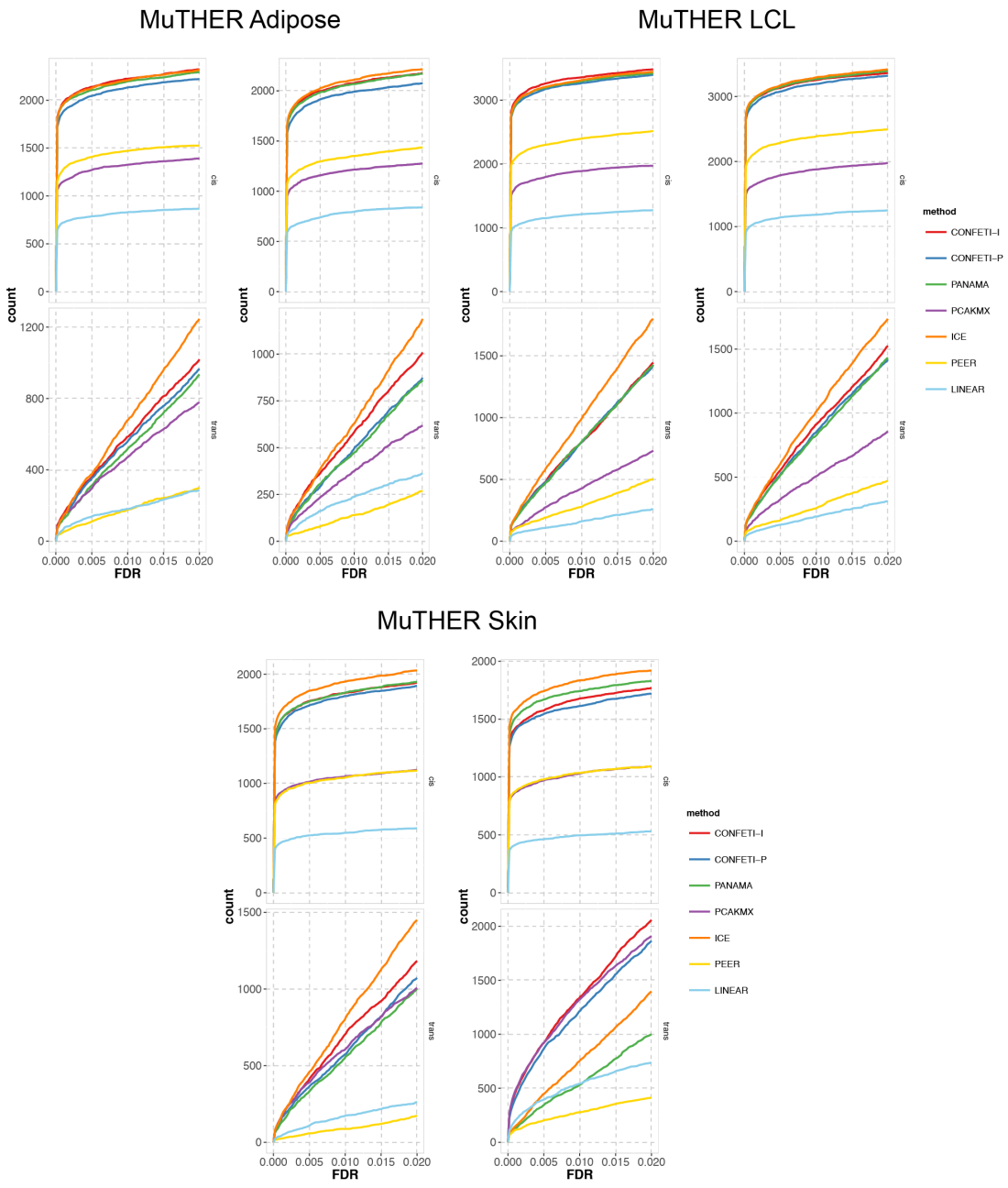


Figure B.3: Significant eQTL discovered in MuTHER datasets for varying FDR thresholds

Plots showing the counts of *cis*- and *trans*-eQTL versus a varying threshold of FDR for each of the methods applied to every dataset.

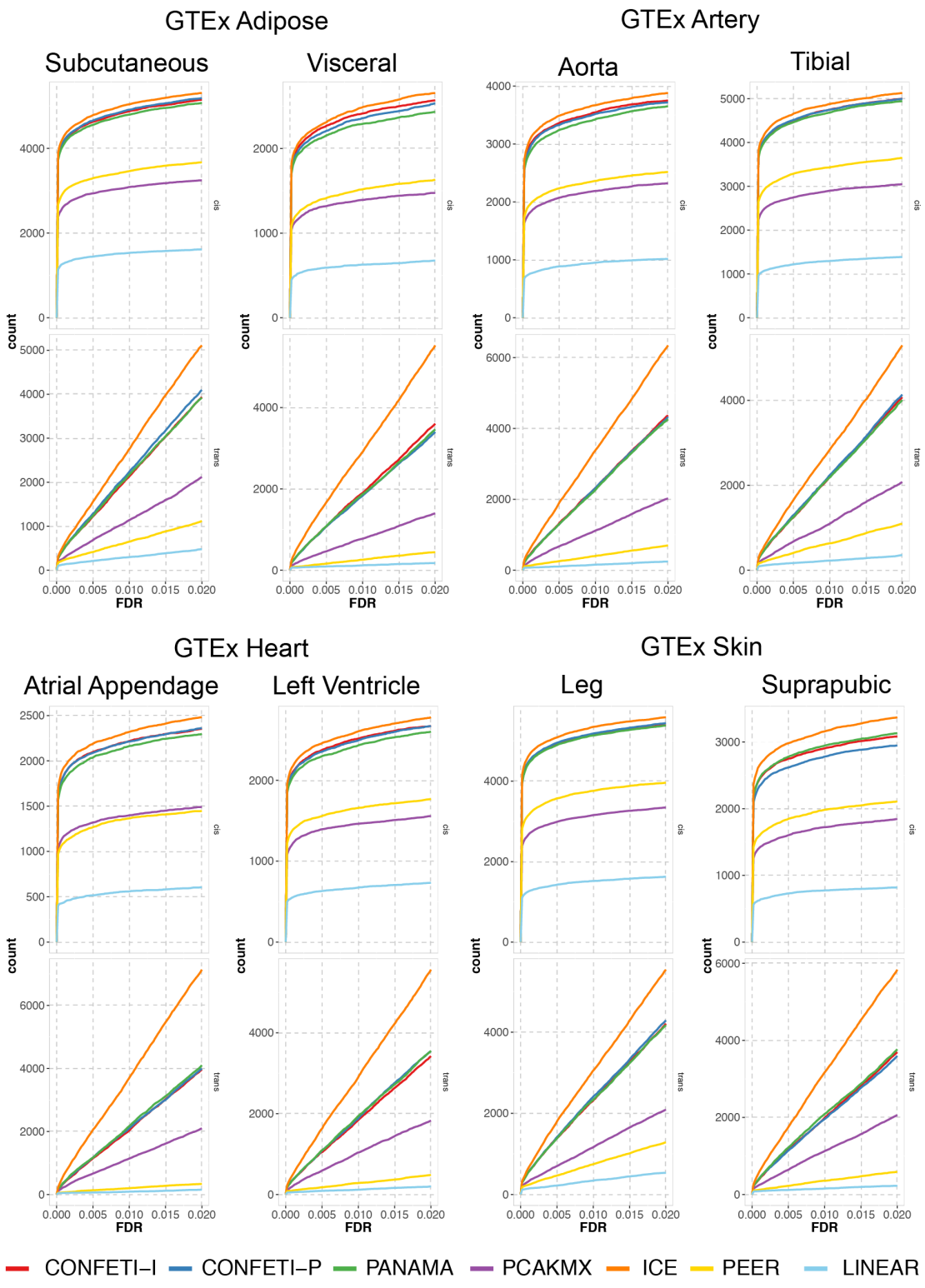


Figure B.4: Significant eQTL discovered in GTEx datasets for varying FDR thresholds.

Plots showing the counts of *cis*- and *trans*-eQTL versus a varying threshold of FDR for each of the methods applied to every dataset.

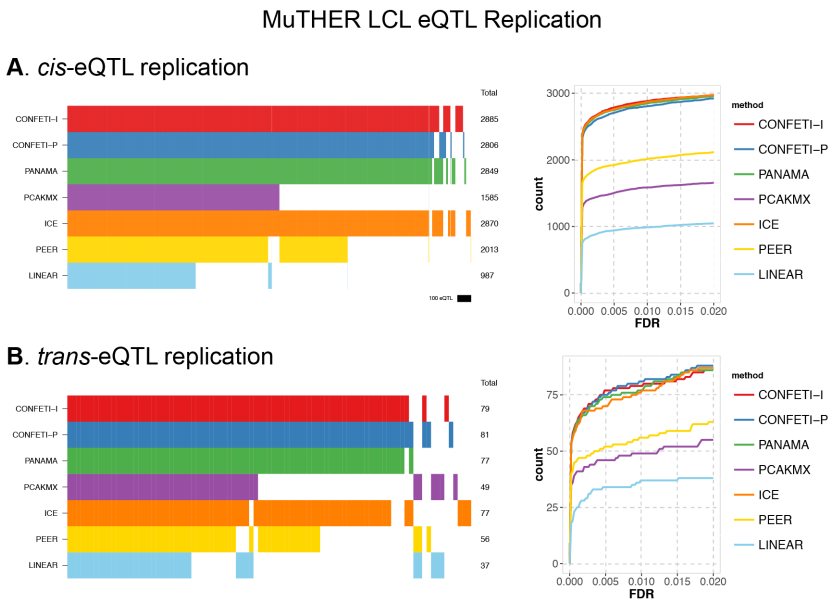


Figure B.5: Replicating eQTL in the MuTHER LCL Subsets.

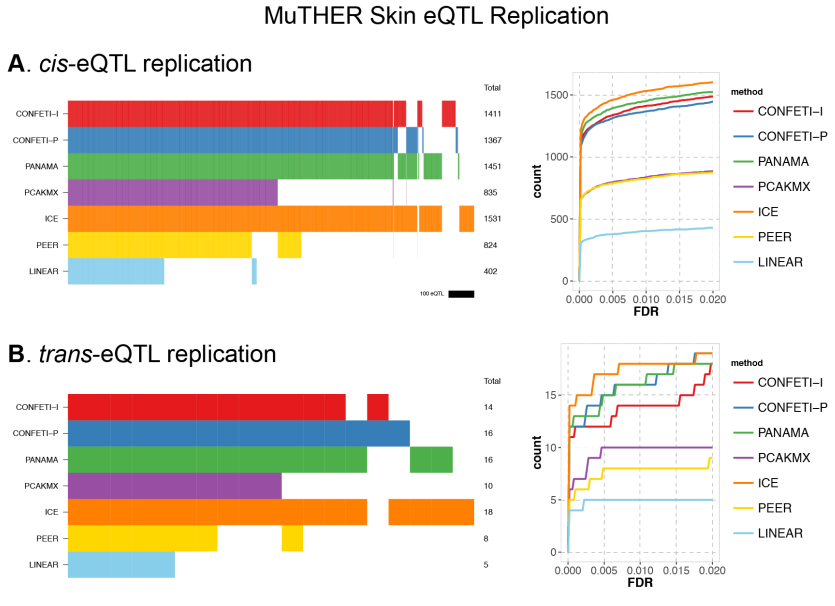


Figure B.6: Replicating eQTL in the MuTHER Skin Subsets.

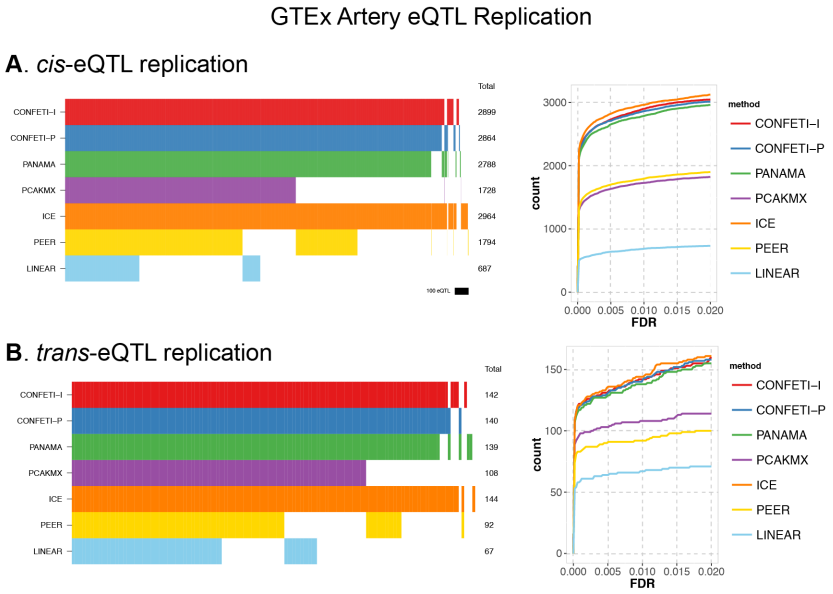


Figure B.7: **Replicating eQTL between GTEx Heart Aorta and Tibial.**

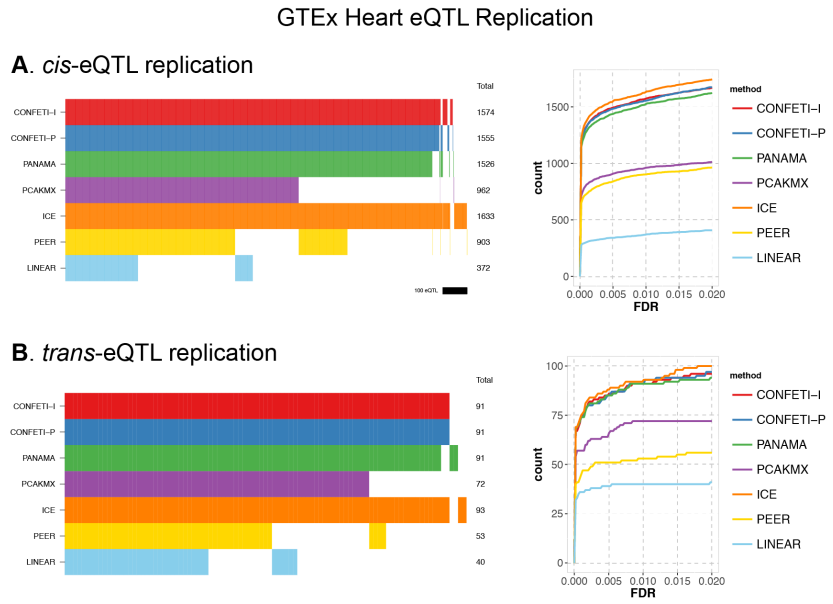


Figure B.8: Replicating eQTL between GTEx Heart Atrial Appendage and Left Ventricle.

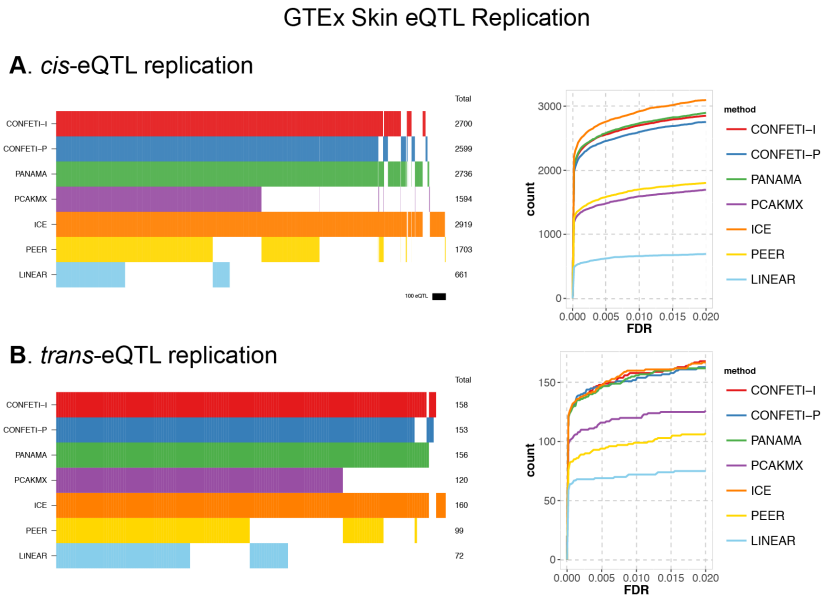


Figure B.9: Replicating eQTL between GTEx Skin Leg and Skin Suprapubic.

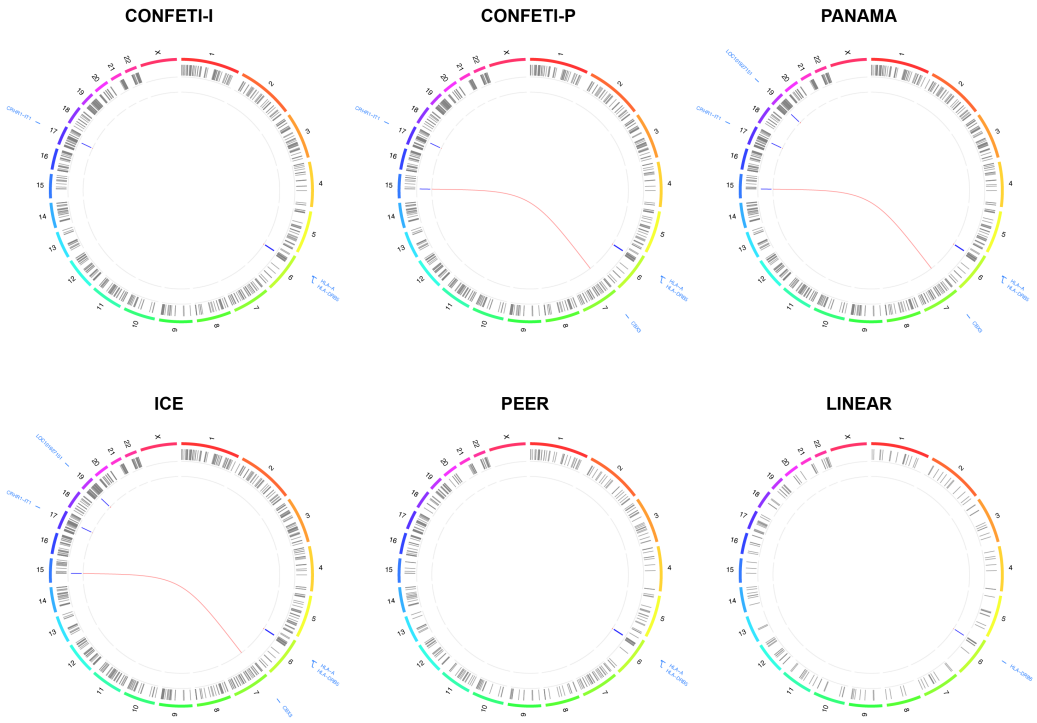


Figure B.10: Circos plots of replicating eQTL across all MuTHER datasets identified by each method.

Chromosomes are plotted in the outermost circles with replicating *cis*-eQTL shown in gray bands within the next layer, and replicating *trans*-eQTL as blue bands in the innermost layer where red lines connect each *trans*-eQTL to the associated gene with gene annotations labeled in blue outside the circle.

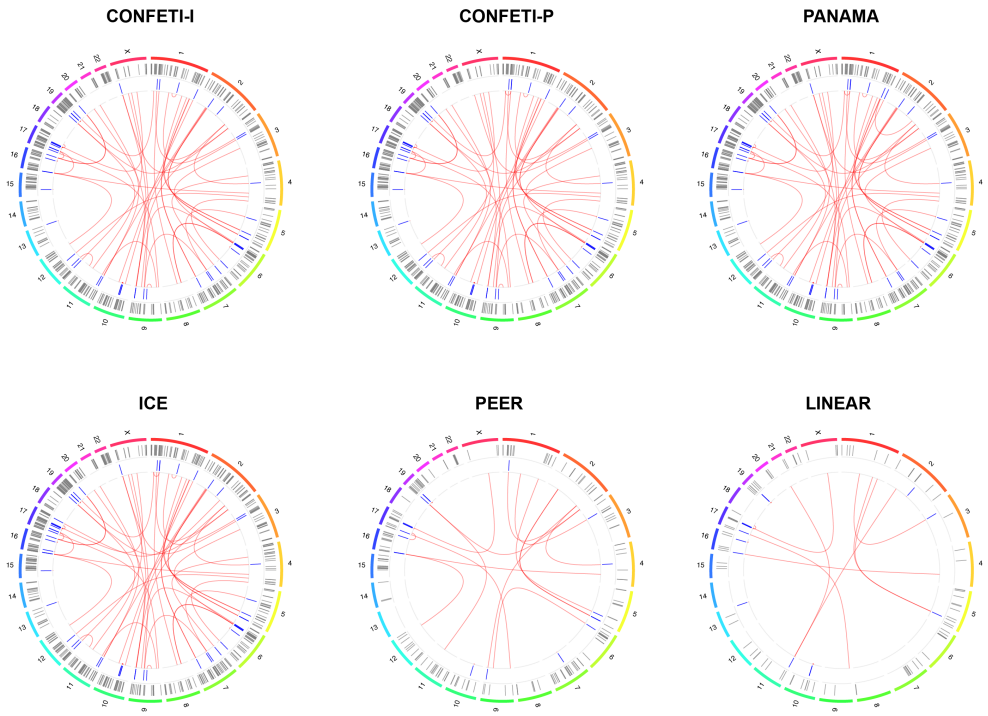


Figure B.11: Circos plots of replicating eQTL across all GTEx datasets identified by each method.

Chromosomes are plotted in the outermost circles with replicating *cis*-eQTL shown in gray bands within the next layer, and replicating *trans*-eQTL as blue bands in the innermost layer where red lines connect each *trans*-eQTL to the associated gene. Gene annotations were excluded due to space limits.

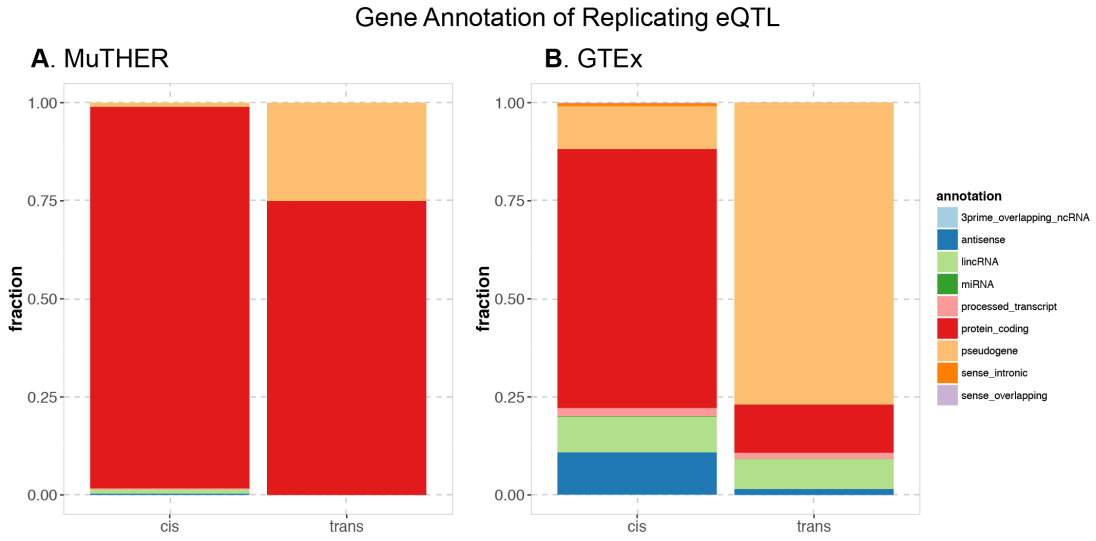


Figure B.12: Gene annotations for replicating eQTL.

Bar plots showing the relative fraction of each gene annotation category in all replicating *cis* and *trans* eQTL identified in **A.** MuTHER and **B.** GTEx.

BIBLIOGRAPHY

[1] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 95–108, 2005.

[2] J. C. Barrett and L. R. Cardon, "Evaluating coverage of genome-wide association studies," *Nature genetics*, vol. 38, no. 6, pp. 659–662, 2006.

[3] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, *et al.*, "Complement factor h polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[4] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, *et al.*, "The nhgri gwas catalog, a curated resource of snp-trait associations," *Nucleic acids research*, vol. 42, no. D1, pp. D1001–D1006, 2014.

[5] B. E. Stranger, E. A. Stahl, and T. Raj, "Progress and promise of genome-wide association studies for human complex trait genetics," *Genetics*, vol. 187, no. 2, pp. 367–383, 2011.

[6] A. L. Price, C. C. Spencer, and P. Donnelly, "Progress and promise in understanding the genetic basis of common diseases," in *Proceedings of the Royal Society B*, vol. 282, p. 20151684, The Royal Society, 2015.

[7] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, *et al.*, "Systematic localization of common disease-associated variation in regulatory dna," *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[8] F. W. Albert and L. Kruglyak, "The role of regulatory variation in complex traits and disease," *Nature Reviews Genetics*, vol. 16, no. 4, pp. 197–212, 2015.

[9] E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, *et al.*, "Genetics of gene expression surveyed in maize, mouse and man," *Nature*, vol. 422, no. 6929, pp. 297–302, 2003.

[10] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung, "Genetic analysis of genome-wide variation in human gene expression," *Nature*, vol. 430, no. 7001, pp. 743–747, 2004.

[11] V. G. Cheung, R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick, "Mapping determinants of human gene expression by regional and genome-wide association," *Nature*, vol. 437, no. 7063, pp. 1365–1369, 2005.

[12] S. Doss, E. E. Schadt, T. A. Drake, and A. J. Lusis, "Cis-acting expression quantitative trait loci in mice," *Genome research*, vol. 15, no. 5, pp. 681–691, 2005.

[13] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, *et al.*, "Relative impact of nucleotide and copy number variation on gene expression phenotypes," *Science*, vol. 315, no. 5813, pp. 848–853, 2007.

[14] H. H. Göring, J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth, S. A. Cole, J. B. Jowett, L. J. Abraham, D. L. Rainwater, A. G. Comuzzie, *et al.*, "Discovery of expression qtls using large-scale transcriptional profiling in human lymphocytes," *Nature genetics*, vol. 39, no. 10, pp. 1208–1216, 2007.

[15] J.-B. Veyrieras, S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard, "High-resolution mapping of expression-qtls yields insight into human gene regulation," *PLoS Genet*, vol. 4, no. 10, p. e1000214, 2008.

[16] E. L. Heinzen, D. Ge, K. D. Cronin, J. M. Maia, K. V. Shianna, W. N. Gabriel, K. A. Welsh-Bohmer, C. M. Hulette, T. N. Denny, and D. B. Goldstein, "Tissue-specific genetic control of splicing: implications for the study of complex traits," *PLoS Biol*, vol. 6, no. 12, p. e1000001, 2008.

[17] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard, "Understanding mechanisms underlying human gene expression variation with rna sequencing," *Nature*, vol. 464, no. 7289, pp. 768–772, 2010.

[18] E. Grundberg, K. S. Small, . K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett, *et al.*, "Mapping cis-and trans-regulatory effects across multiple tissues in twins," *Nature genetics*, vol. 44, no. 10, pp. 1084–1089, 2012.

[19] D. Mehta, K. Heim, C. Herder, M. Carstensen, G. Eckstein, C. Schurmann, G. Homuth, M. Nauck, U. Völker, M. Roden, *et al.*, "Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood," *European Journal of Human Genetics*, vol. 21, no. 1, pp. 48–54, 2013.

[20] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. ACt Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, *et al.*, "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, vol. 501, no. 7468, pp. 506–511, 2013.

[21] A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, *et al.*, "Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals," *Genome research*, vol. 24, no. 1, pp. 14–24, 2014.

[22] G. Consortium *et al.*, "The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, 2015.

[23] M. F. Moffatt, M. Kabesch, L. Liang, A. L. Dixon, D. Strachan, S. Heath, M. Depner, A. von Berg, A. Bufe, E. Rietschel, *et al.*, "Genetic variants regulating ormdl3 expression contribute to the risk of childhood asthma," *Nature*, vol. 448, no. 7152, pp. 470–473, 2007.

[24] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox, "Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas," *PLoS Genet*, vol. 6, no. 4, p. e1000888, 2010.

[25] K. Musunuru, A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt, K. V. Sachs, X. Li, H. Li, N. Kuperwasser, V. M. Ruda, *et al.*, "From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus," *Nature*, vol. 466, no. 7307, pp. 714–719, 2010.

[26] A. C. Nica, S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso, and E. T. Dermitzakis, "Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations," *PLoS Genet*, vol. 6, no. 4, p. e1000895, 2010.

[27] P. C. Dubois, G. Trynka, L. Franke, K. A. Hunt, J. Romanos, A. Curtotti, A. Zhernakova, G. A. Heap, R. Ádány, A. Aromaa, *et al.*, "Multiple common variants for celiac disease influencing immune gene expression," *Nature genetics*, vol. 42, no. 4, pp. 295–302, 2010.

[28] H. H. Nguyen, R. Takata, S. Akamatsu, D. Shigemizu, T. Tsunoda, M. Furihata, A. Takahashi, M. Kubo, N. Kamatani, O. Ogawa, *et al.*, "Irx4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin d receptor, conferring prostate cancer susceptibility," *Human molecular genetics*, p. dds025, 2012.

[29] F. Zou, H. S. Chai, C. S. Younkin, M. Allen, J. Crook, V. S. Pankratz, M. M. Carrasquillo, C. N. Rowley, A. A. Nair, S. Middha, *et al.*, "Brain expression genome-wide association study (egwas) identifies human disease-associated variants," *PLoS Genet*, vol. 8, no. 6, p. e1002707, 2012.

[30] H.-J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, *et al.*, "Systematic identification of trans eqtls as putative drivers of known disease associations," *Nature genetics*, vol. 45, no. 10, pp. 1238–1243, 2013.

[31] C. L. Miller, D. R. Anderson, R. K. Kundu, A. Raiesdana, S. T. Nürnberg, R. Diaz, K. Cheng, N. J. Leeper, C.-H. Chen, I.-S. Chang, *et al.*, "Disease-related growth factor and embryonic signaling pathways modulate an enhancer of tcf21 expression at the 6q23. 2 coronary heart disease locus," *PLoS Genet*, vol. 9, no. 7, p. e1003652, 2013.

[32] M. Lamontagne, C. Couture, D. S. Postma, W. Timens, D. D. Sin, P. D. Pare, J. C. Hogg, D. Nickle, M. Laviolette, and Y. Bosse, "Refining susceptibility loci of chronic obstructive pulmonary disease with lung eqtls," *PLoS One*, vol. 8, no. 7, p. e70220, 2013.

[33] V. Kumar, H.-J. Westra, J. Karjalainen, D. V. Zhernakova, T. Esko, B. Hrdlickova, R. Almeida, A. Zhernakova, E. Reinmaa, U. Võsa, *et al.*, "Human disease-associated genetic variation impacts large intergenic non-coding rna expression," *PLoS Genet*, vol. 9, no. 1, p. e1003201, 2013.

[34] T. Raj, K. Rothamel, S. Mostafavi, C. Ye, M. N. Lee, J. M. Replogle, T. Feng, M. Lee, N. Asinovski, I. Frohlich, *et al.*, "Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes," *Science*, vol. 344, no. 6183, pp. 519–523, 2014.

[35] T. Singh, A. P. Levine, P. J. Smith, A. M. Smith, A. W. Segal, and J. C. Barrett, "Characterization of expression quantitative trait loci in the human colon," *Inflammatory bowel diseases*, vol. 21, no. 2, p. 251, 2015.

[36] E. T. Dermitzakis, "From gene expression to disease risk," *Nature genetics*, vol. 40, no. 5, pp. 492–493, 2008.

[37] Y. Gilad, S. A. Rifkin, and J. K. Pritchard, "Revealing the architecture of gene regulation: the promise of eqtl studies," *Trends in genetics*, vol. 24, no. 8, pp. 408–415, 2008.

[38] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, "Mapping complex disease traits with global gene expression," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 184–194, 2009.

[39] K. Fransen, M. C. Visschedijk, S. van Sommeren, J. Y. Fu, L. Franke, E. A. Festen, P. C. Stokkers, A. A. van Bodegraven, J. B. A. Crusius, D. W. Hommes, *et al.*, "Analysis of snps with an effect on gene expression identifies ube2l3 and bcl3 as potential new risk genes for crohn's disease," *Human molecular genetics*, vol. 19, no. 17, pp. 3482–3488, 2010.

[40] H. Zhong, J. Beaulaurier, P. Y. Lum, C. Molony, X. Yang, D. J. MacNeil, D. T. Weingarth, B. Zhang, D. Greenawalt, R. Dobrin, *et al.*, "Liver and adipose expression associated snps are enriched for association to type 2 diabetes," *PLoS Genet*, vol. 6, no. 5, p. e1000932, 2010.

[41] S. B. Montgomery and E. T. Dermitzakis, "From expression qtls to personalized transcriptomics," *Nature Reviews Genetics*, vol. 12, no. 4, pp. 277–282, 2011.

[42] H. P. Kang, A. A. Morgan, R. Chen, E. E. Schadt, and A. J. Butte, "Coanalysis of gwas with eqtls reveals disease-tissue associations," *AMIA Summits on Translational Science proceedings*, vol. 2012, p. 35, 2012.

[43] A. L. Richards, L. Jones, V. Moskvina, G. Kirov, P. V. Gejman, D. F. Levinson, A. R. Sanders, S. Purcell, P. M. Visscher, N. Craddock, *et al.*, "Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain," *Molecular psychiatry*, vol. 17, no. 2, pp. 193–201, 2012.

[44] S. L. Edwards, J. Beesley, J. D. French, and A. M. Dunning, "Beyond gwass: illuminating the dark road from association to function," *The American Journal of Human Genetics*, vol. 93, no. 5, pp. 779–797, 2013.

[45] X. He, C. K. Fuller, Y. Song, Q. Meng, B. Zhang, X. Yang, and H. Li, "Sherlock: detecting gene-disease associations by matching patterns of expression qtl and gwas," *The American Journal of Human Genetics*, vol. 92, no. 5, pp. 667–680, 2013.

[46] E. R. Gamazon, W. Zhang, A. Konkashbaev, S. Duan, E. O. Kistner, D. L. Nicolae, M. E. Dolan, and N. J. Cox, "Scan: Snp and copy number annotation," *Bioinformatics*, vol. 26, no. 2, pp. 259–262, 2010.

[47] H. Zhong, X. Yang, L. M. Kaplan, C. Molony, and E. E. Schadt, "Integrating pathway analysis and genetics of gene expression for genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, no. 4, pp. 581–591, 2010.

[48] M. Civelek and A. J. Lusis, "Systems genetics approaches to understand complex traits," *Nature Reviews Genetics*, vol. 15, no. 1, pp. 34–48, 2014.

[49] K. A. Williams, M. Lee, Y. Hu, J. Andreas, S. J. Patel, S. Zhang, P. Chines, A. Elkahloun, S. Chandrasekharappa, J. S. Gutkind, *et al.*, "A systems genetics approach identifies cxcl14, itgax, and lpcat2 as novel aggressive prostate cancer susceptibility genes," *PLoS Genet*, vol. 10, no. 11, p. e1004809, 2014.

[50] M. R. Johnson, J. Behmoaras, L. Bottolo, M. L. Krishnan, K. Pernhorst, P. L. M. Santoscoy, T. Rossetti, D. Speed, P. K. Srivastava, M. Chadeau-Hyam, *et al.*, "Systems genetics identifies sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus," *Nature communications*, vol. 6, 2015.

[51] J. Wang, M. C. J. Ma, A. K. Mennie, J. M. Pettus, Y. Xu, L. Lin, M. G. Traxler, J. Jakoubek, S. S. Atanur, T. J. Aitman, *et al.*, "Systems biology with high-throughput sequencing reveals genetic mechanisms underlying the metabolic syndrome in the lyon hypertensive rat," *Circulation: Cardiovascular Genetics*, vol. 8, no. 2, pp. 316–326, 2015.

[52] J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt, "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," *Nature genetics*, vol. 40, no. 7, pp. 854–861, 2008.

[53] R. H. Blair, D. J. Kliebenstein, and G. A. Churchill, "What can causal networks tell us about metabolic pathways?," *PLoS Comput Biol*, vol. 8, no. 4, p. e1002458, 2012.

[54] V.-P. Mäkinen, M. Civelek, Q. Meng, B. Zhang, J. Zhu, C. Levian, T. Huan, A. V. Segrè, S. Ghosh, J. Vivar, *et al.*, "Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease," *PLoS Genet*, vol. 10, no. 7, p. e1004502, 2014.

[55] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L. Svenson, G. A. Churchill, and S. P. Gygi, "Defining the consequences of genetic variation on a proteome-wide scale," *Nature*, vol. 534, no. 7608, pp. 500–505, 2016.

[56] B. E. Stranger, M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch, R. Lyle, S. Hunt, B. Kahl, S. E. Antonarakis, S. Tavaré, *et al.*, "Genome-wide associations of gene expression variation in humans," *PLoS Genet*, vol. 1, no. 6, p. e78, 2005.

[57] A. A. Pai, J. K. Pritchard, and Y. Gilad, "The genetic and mechanistic basis for variation in gene regulation," *PLoS Genet*, vol. 11, no. 1, p. e1004857, 2015.

[58] E. Petretto, J. Mangion, N. J. Dickens, S. A. Cook, M. K. Kumaran, H. Lu, J. Fischer, H. Maatz, V. Kren, M. Pravenec, *et al.*, "Heritability and tissue specificity of expression quantitative trait loci," *PLoS Genet*, vol. 2, no. 10, p. e172, 2006.

[59] A. L. Price, A. Helgason, G. Thorleifsson, S. A. McCarroll, A. Kong, and K. Stefansson, "Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals," *PLoS Genet*, vol. 7, no. 2, p. e1001317, 2011.

[60] B. P. Fairfax, S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight, "Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of hla alleles," *Nature genetics*, vol. 44, no. 5, pp. 502–510, 2012.

[61] D. J. Gaffney, "Global properties and functional complexity of human gene regulatory variation," *PLoS Genet*, vol. 9, no. 5, p. e1003501, 2013.

[62] B. P. Fairfax, P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, C. McGee, *et al.*, "Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression," *Science*, vol. 343, no. 6175, p. 1246949, 2014.

[63] Y. Chen, J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts, *et al.*, "Variations in dna elucidate molecular networks that cause disease," *Nature*, vol. 452, no. 7186, pp. 429–435, 2008.

[64] H. Kirsten, H. Al-Hasani, L. Holdt, A. Gross, F. Beutner, K. Krohn, K. Horn, P. Ahnert, R. Burkhardt, K. Reiche, *et al.*, "Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eqtls and corroborates the regulatory relevance of non-protein coding loci," *Human molecular genetics*, vol. 24, no. 16, pp. 4746–4763, 2015.

[65] M. Consortium *et al.*, "Identification of an imprinted master trans regulator at the klf14 locus related to multiple metabolic phenotypes," *Nature genetics*, vol. 43, no. 6, pp. 561–564, 2011.

[66] H. M. Kang, C. Ye, and E. Eskin, "Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots," *Genetics*, vol. 180, no. 4, pp. 1909–1925, 2008.

[67] A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E. E. Schadt, T. A. Drake, A. J. Lusis, *et al.*, "Integrating genetic and network analysis to characterize genes related to mouse weight," *PLoS Genet*, vol. 2, no. 8, p. e130, 2006.

[68] C. Wu, D. L. Delano, N. Mitro, S. V. Su, J. Janes, P. McClurg, S. Batalov, G. L. Welch, J. Zhang, A. P. Orth, *et al.*, "Gene set enrichment in eqtl data identifies novel annotations and pathway regulators," *PLoS Genet*, vol. 4, no. 5, p. e1000070, 2008.

[69] B. A. Logsdon and J. Mezey, "Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations," *PLoS Comput Biol*, vol. 6, no. 12, p. e1001014, 2010.

[70] M. Heinig, E. Petretto, C. Wallace, L. Bottolo, M. Rotival, H. Lu, Y. Li, R. Sarwar, S. R. Langley, A. Bauerfeind, *et al.*, "A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk," *Nature*, vol. 467, no. 7314, pp. 460–464, 2010.

[71] A. Aterido, C. Palacio, S. Marsal, G. Ávila, and A. Julià, "Novel insights into the regulatory architecture of cd4+ t cells in rheumatoid arthritis," *PloS one*, vol. 9, no. 6, p. e100690, 2014.

[72] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak, "Genetic dissection of transcriptional regulation in budding yeast," *Science*, vol. 296, no. 5568, pp. 752–755, 2002.

[73] R. B. Brem and L. Kruglyak, "The landscape of genetic complexity across

5,700 gene expression traits in yeast," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 5, pp. 1572–1577, 2005.

[74] E. J. Foss, D. Radulovic, S. A. Shaffer, D. M. Ruderfer, A. Bedalov, D. R. Goodlett, and L. Kruglyak, "Genetic basis of proteome variation in yeast," *Nature genetics*, vol. 39, no. 11, pp. 1369–1375, 2007.

[75] A. van Nas, L. Ingram-Drake, J. S. Sinsheimer, S. S. Wang, E. E. Schadt, T. Drake, and A. J. Lusis, "Expression quantitative trait loci: replication, tissue-and sex-specificity in mice," *Genetics*, vol. 185, no. 3, pp. 1059–1068, 2010.

[76] O. Stegle, L. Parts, R. Durbin, and J. Winn, "A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies," *PLoS Comput Biol*, vol. 6, no. 5, p. e1000770, 2010.

[77] N. Fusi, O. Stegle, and N. D. Lawrence, "Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies," *PLoS Comput Biol*, vol. 8, no. 1, p. e1002330, 2012.

[78] D. J. Gaffney, J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard, "Dissecting the regulatory architecture of gene expression qtls," *Genome biology*, vol. 13, no. 1, p. 1, 2012.

[79] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, no. 10, pp. 781–791, 2006.

[80] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[81] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[82] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, "Efficient control of population structure in model organism association mapping," *Genetics*, vol. 178, no. 3, pp. 1709–1723, 2008.

[83] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, E. Eskin, *et al.*, "Variance component model to account for sample structure in genome-wide association studies," *Nature genetics*, vol. 42, no. 4, pp. 348–354, 2010.

[84] Z. Zhang, E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, *et al.*, "Mixed linear model approach adapted for genome-wide association studies," *Nature genetics*, vol. 42, no. 4, pp. 355–360, 2010.

[85] A. Raj, M. Stephens, and J. K. Pritchard, "faststructure: variational inference of population structure in large snp data sets," *Genetics*, vol. 197, no. 2, pp. 573–589, 2014.

[86] T. L. Fare, E. M. Coffey, H. Dai, Y. D. He, D. A. Kessler, K. A. Kilian, J. E. Koch, E. LeProust, M. J. Marton, M. R. Meyer, *et al.*, "Effects of atmospheric ozone on microarray data quality," *Analytical chemistry*, vol. 75, no. 17, pp. 4672–4675, 2003.

[87] S. Li, P. P. Łabaj, P. Zumbo, P. Sykacek, W. Shi, L. Shi, J. Phan, P.-Y. Wu, M. Wang, C. Wang, *et al.*, "Detecting and correcting systematic variation in large-scale rna sequencing data," *Nature biotechnology*, vol. 32, no. 9, pp. 888–895, 2014.

[88] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genet*, vol. 3, no. 9, p. e161, 2007.

[89] J. Listgarten, C. Kadie, E. E. Schadt, and D. Heckerman, "Correction for hidden confounders in the genetic analysis of gene expression," *Proceedings of the National Academy of Sciences*, vol. 107, no. 38, pp. 16465–16470, 2010.

[90] A. E. Teschendorff, J. Zhuang, and M. Widschwendter, "Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies," *Bioinformatics*, vol. 27, no. 11, pp. 1496–1505, 2011.

[91] C. Gao, N. L. Tignor, J. Salit, Y. Strulovici-Barel, N. R. Hackett, R. G. Crystal, and J. G. Mezey, "Heft: eqtl analysis of many thousands of expressed genes while simultaneously controlling for hidden factors," *Bioinformatics*, p. btt690, 2013.

[92] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "Fast linear mixed models for genome-wide association studies," *Nature methods*, vol. 8, no. 10, pp. 833–835, 2011.

[93] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[94] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010.

[95] A. E. Teschendorff, M. Journée, P. A. Absil, R. Sepulchre, and C. Caldas, "Elucidating the altered transcriptional programs in breast cancer using independent component analysis," *PLoS Comput Biol*, vol. 3, no. 8, p. e161, 2007.

[96] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, no. 1, pp. 51–60, 2002.

[97] S.-I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome biology*, vol. 4, no. 11, p. 1, 2003.

[98] J. M. Engreitz, B. J. Daigle, J. J. Marshall, and R. B. Altman, "Independent component analysis: mining microarray data for fundamental human gene expression modules," *Journal of biomedical informatics*, vol. 43, no. 6, pp. 932–944, 2010.

[99] C. H. Bang-Berthelsen, L. Pedersen, T. Fløyel, P. H. Hagedorn, T. Gylvin, and F. Pociot, "Independent component and pathway-based analysis of mirna-regulated gene expression in a model of type 1 diabetes," *BMC genomics*, vol. 12, no. 1, p. 97, 2011.

[100] A. Biton, I. Bernard-Pierrot, Y. Lou, C. Krucker, E. Chapeaublanc, C. Rubio-Pérez, N. López-Bigas, A. Kamoun, Y. Neuzillet, P. Gestraud, *et al.*, "Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes," *Cell reports*, vol. 9, no. 4, pp. 1235–1245, 2014.

[101] S. Biswas, J. D. Storey, and J. M. Akey, "Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 1, 2008.

[102] M. Rotival, T. Zeller, P. S. Wild, S. Maouche, S. Szymczak, A. Schillert, R. Castagné, A. Deiseroth, C. Proust, J. Brocheton, *et al.*, "Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans," *PLoS Genet*, vol. 7, no. 12, p. e1002367, 2011.

[103] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, vol. 46. John Wiley & Sons, 2004.

[104] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[105] J.-F. Cardoso, "High-order contrasts for independent component analysis," *Neural computation*, vol. 11, no. 1, pp. 157–192, 1999.

[106] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of machine learning research*, vol. 3, no. Jul, pp. 1–48, 2002.

[107] E. G. Learned-Miller and W. F. John III, "Ica using spacings estimates of entropy," *Journal of Machine Learning Research*, vol. 4, no. Dec, pp. 1271–1295, 2003.

[108] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[109] A. Hyvrinen, "New approximations of differential entropy for independent component analysis and projection pursuit," in *Proceedings of the 1997 conference on Advances in neural information processing systems*, vol. 10, pp. 273–279, 1998.

[110] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca, *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*, 2012. Technical Report No. 597.

[111] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[112] N. Dhingra, A. Shemer, J. C. da Rosa, M. Rozenblit, J. Fuentes-Duculan, J. K. Gittler, R. Finney, T. Czarnowicki, X. Zheng, H. Xu, *et al.*, "Molecular profiling of contact dermatitis skin identifies allergen-dependent dif-

ferences in immune response," *Journal of Allergy and Clinical Immunology*, vol. 134, no. 2, pp. 362–372, 2014.

[113] P. AC't Hoen, M. R. Friedländer, J. Almlöf, M. Sammeth, I. Pulyakhina, S. Y. Anvar, J. F. Laros, H. P. Buermans, O. Karlberg, M. Brännvall, *et al.*, "Reproducibility of high-throughput mrna and small rna sequencing across laboratories," *Nature biotechnology*, vol. 31, no. 11, pp. 1015–1022, 2013.

[114] C. Yang, L. Wang, S. Zhang, and H. Zhao, "Accounting for non-genetic factors by low-rank representation and sparse regression for eqtl mapping," *Bioinformatics*, vol. 29, no. 8, pp. 1026–1034, 2013.

[115] J. W. J. Joo, J. H. Sul, B. Han, C. Ye, and E. Eskin, "Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies," *Genome biology*, vol. 15, no. 4, p. 1, 2014.

[116] S. Mostafavi, A. Battle, X. Zhu, A. E. Urban, D. Levinson, S. B. Montgomery, and D. Koller, "Normalizing rna-sequencing data by modeling hidden covariates with prior knowledge," *PLoS One*, vol. 8, no. 7, p. e68141, 2013.

[117] A. Goldinger, A. K. Henders, A. F. McRae, N. G. Martin, G. Gibson, G. W. Montgomery, P. M. Visscher, and J. E. Powell, "Genetic and nongenetic variation revealed for the principal components of human gene expression," *Genetics*, vol. 195, no. 3, pp. 1117–1128, 2013.

[118] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[119] O. Stegle, A. Kannan, R. Durbin, and J. Winn, "Accounting for non-genetic factors improves the power of eqtl studies," in *Annual International Conference on Research in Computational Molecular Biology*, pp. 411–422, Springer, 2008.

[120] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin, "Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses," *Nature protocols*, vol. 7, no. 3, pp. 500–507, 2012.

[121] C. Lippert, F. P. Casale, B. Rakitsch, and O. Stegle, "Limix: genetic analysis of multiple traits," *BioRxiv*, p. 003905, 2014.

[122] J. L. Marchini, C. Heaton, and B. D. Ripley, *fastICA: FastICA Algorithms to perform ICA and Projection Pursuit*, 2013. R package version 1.2-0.

[123] A. Frigyesi, S. Veerla, D. Lindgren, and M. Höglund, "Independent component analysis reveals new and biologically significant structures in micro array data," *BMC bioinformatics*, vol. 7, no. 1, p. 1, 2006.

[124] M. N. Lee, C. Ye, A.-C. Villani, T. Raj, W. Li, T. M. Eisenhaure, S. H. Imboywa, P. I. Chipendo, F. A. Ran, K. Slowikowski, *et al.*, "Common genetic variants modulate pathogen-sensing responses in human dendritic cells," *Science*, vol. 343, no. 6175, p. 1246980, 2014.

[125] K. Lawrenson, Q. Li, S. Kar, J.-H. Seo, J. Tyrer, T. J. Spindler, J. Lee, Y. Chen, A. Karst, R. Drapkin, *et al.*, "Cis-eqtl analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer," *Nature communications*, vol. 6, 2015.

[126] E. N. Smith and L. Kruglyak, "Gene–environment interaction in yeast gene expression," *PLoS Biol*, vol. 6, no. 4, p. e83, 2008.

[127] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[128] W. J. Kent, "BLAT-the blast-like alignment tool," *Genome research*, vol. 12, no. 4, pp. 656–664, 2002.